Detecting Sexually Provocative Images

Debashis Ganguly* Mohammad H. Mofrad* Adriana Kovashka
Department of Computer Science
University of Pittsburgh
{debashis, hasanzadeh, kovashka}@cs.pitt.edu

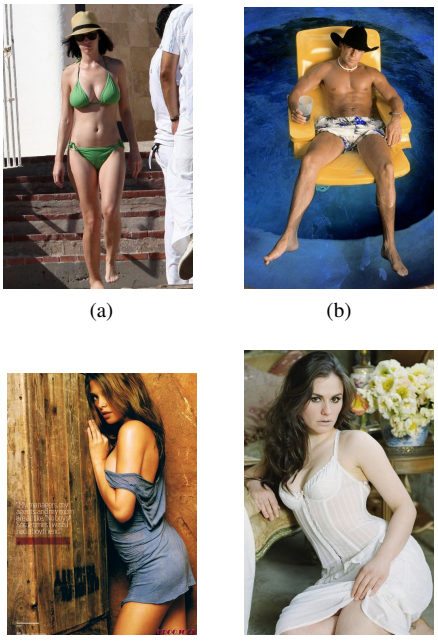
Abstract

While the abundance of visual content available on the Internet, and the easy access to such content by all users allows us to find relevant content quickly, it also poses challenges. For example, if a parent wants to restrict the visual content which their child can see, this content needs to either be automatically tagged as offensive or not, or a computer vision algorithm needs to be trained to detect offensive content. One type of potentially offensive content is sexually explicit or provocative imagery. An image may be sexually provocative if it portrays nudity, but the sexual innuendo could also be contained in the body posture or facial expression of the human subject shown in the photo. Existing methods simply analyze skin exposure, but fail to capture the hidden intent behind images. Thus, they are unable to capture several important ways in which an image might be sexually provocative, hence offensive to children. We propose to address this problem by extracting a unified feature descriptor constituting the percentage of skin exposure, the body posture of the human in the image, and his/her gestures and facial expressions. We learn to predict these cues, then train a hierarchical model which combines them. We show in experiments that this model more accurately detects sexual innuendos behind images.

1. Introduction

There is an overwhelming amount of visual data on the Internet today. For example, each day, 300 million photographs are uploaded to Flickr [1], and over 500 thousand hours of video are uploaded to YouTube [2]. This data allows users on the web to find content that has a wide variety of uses (e.g. for websites, presentations, etc.). However, this abundance poses a challenge for anyone who wishes to automatically flag and filter this content.

Search engines have the ability to filter images for offensive content, but this filtering is not perfect and often allows offensive imagery in the "safe search" results. One type of such offensive content is sexually provocative imagery. Unfortunately, existing methods primarily detect sexual content by analyzing the amount of skin exposure shown in the photograph.



(a) (b) (c) (d)

Figure 1: Sexually provocative content consists of more than skin exposure. For example, images (a) and (b) show a significant amount of skin exposure, but do not appears sexually provocative. In contrast, (c) and (d) show a smaller amount of skin, but are more sexually suggestive.

However, there is a difference between what an image apparently portrays, and what is expressed by the body-language and expressions of the subject. Consider the images in Figure 1. Images (a) and (b) show subjects who are almost nude, but most humans will agree that these images contain no sexual intent. On the other hand, images (c) and

*These authors contributed equally to the work.

(d) do not contain nude subjects, but they clearly show sexual intent.

Intentions play a core role in communication and persuasion. In portrait photography, the photographer and subject have some idea of how the audience should perceive the subject, e.g., as having certain qualities. In pornographic visual imagery, the intent of the subject is to influence the interpretation of the image by a viewer. Photographers identify these cues and essentially instruct the subject to pose for such specific impression to influence the judgement of the viewers. In this paper, we examine such cues behind images to characterize **sexual intentions** behind popular celebrity images. We exploit 17 types of attributes composed from **facial expressions, postures, and gestures**. We also incorporate a measure of the **amount of skin exposed**, and a feature dimension which captures the **scene context** information in the photo. These features are in turn used to predict the **mood and emotion** of the subject that can help deduce the **sexual intent** of the overall image context. In summary this work has following major contributions:

1. We define the novel problem of inferring the sexual intentions behind celebrity images.

2. We develop a hierarchical approach to detect sexual intention. First, using automatically extracted computer vision features, we predict the body pose and facial expressions of the subject. We also capture the skin exposure of the subject, and the image background. We refer to these cues collectively as "attributes". We use these to infer the mood of the subject, and in turn use the latter to predict sexual intent or lack thereof.

3. We present a dataset of 1,146 images of 203 celebrities annotated with visual cues to identify sexual intentions. The dataset and collected annotations is made available for public download[1].

The remainder of the paper is organized as follows. Section 2 presents the related work and shows how the current work differs from the existing ones. Section 3 describes the hierarchical method along with attribute and mood annotations used for this method. Section 4 presents the dataset we collected, and our experimental results. Section 5 concludes the paper.

2. Related Work

Skin detection algorithms are among the most basic methods for filtering sexual contents in images. Jiao et al. [10] use the color coherence vector and color histogram of images to train a classifier for detecting nudity in images. In another work [6], the skin color model is applied to the binary classification problem of adult image detection. The authors employed color and texture filters to extract the skin regions and feed them to a classifier. In [20], a skin pixel detection algorithm is employed to first binarize the image into skin regions and non-skin regions and then to refine these regions with erosion and dilation morphological operators. Finally, they extract the shape features from these regions and then use Adaboost to classify the images into sexual and non-sexual categories. Another learning algorithm based on Adaboost is proposed in [14] where a chromatic distribution matching scheme is used to precisely determine skin chroma distribution. This framework utilizes skin segmentation while effectively combining geometric constraints of naked bodies using Adaboost. Finally, a neural network is used to achieve the final classification results.

In addition to skin detection algorithms, there exists some other sophisticated methods for detecting sexual innuendos. Addressing the problem of adult content detection in images on the Internet, Rowley et al. [16] create a search engine plug-in for adult content filtering. First, they extract skin-dependent features (e.g. skin map and texture) and skin-independent features (e.g. shape, size, entropy, etc.) of a set of 17,300 annotated images. Then, they train an SVM to evaluate the contribution of these features. Their proposed system is able to detect roughly 50% of the adult-content images. Deselaers et al. [5] use the bag of visual words with a task-specific visual vocabulary trained on a dataset of 8,500 images. They identify the interest points in an image and extract image patches around them. Then, they create a visual vocabulary by training a Gaussian mixture model. Finally, the histogram of this visual vocabulary is computed and given to a discriminant classifier to predict a specific class of pornographic images. Hu et al. [7] propose an algorithm for recognizing pornographic web pages using a decision tree. In their text and image fusion algorithm, the Bayes theorem is employed to combine the recognition results from images and texts. Their hybrid approach outperforms the contour-based and skin-region-based classifiers for detecting pornographic images.

The research in [10, 6, 20, 14, 16, 5, 7] focuses on extracting the skin regions and classifying the images for adult contents based on the amount of exposure. These cited works pay little attention to the intention behind the image composition and the goals of the photographic subject with respect to how the photo should be perceived. This is a limitation of existing approaches as they cannot differentiate between "naked" subjects and "sexual" subjects. For example, these schemes would fail to distinguish behind a portrait like "Mona Lisa", a harmless photo of a bodybuilder taken at gym, and a person posing with sexual intention. Our approach aims to identify sexual innuendos behind images beyond the apparent nudity of the subject.

Our research is highly motivated by recent work in in-

[1] https://github.com/DebashisGanguly/SexualIntentDetection

ferring the intent of the photographer in portraits of politicians [11, 12, 8]. While this work does not study sexually provocative imagery, it shares the goal of our work in analyzing images beyond what is obvious. Joo et al. [11] introduce a new problem to the field of computer vision. They put forward the new topic of understanding visual persuasion from mass media images. They propose a hierarchical model that predicts persuasive intents with respect to the qualities of the subjects that are portrayed (e.g. *energetic*, *trustworthy*, etc.). The model learns to predict these intents from "syntactical attributes" of the subject (e.g. *smile*, *waving hands*, etc.). [11] constructed a dataset of 1,124 images of politicians which are labeled with their corresponding intents. Similar to [11], we also adopt a hierarchical framework, but we adapt it to use attributes relevant to our task. In another work, Joo et al. [12] study the face of politicians for inferring social traits (e.g *intelligence*, *honesty*, etc.) and apply them to predict a social event (e.g. *presidential debate*, *U.S. elections*, etc.). Their method classifies the winners of a series of recent U.S. elections with the approximated accuracy of 65%. It also categorizes the politician images into their corresponding political party (i.e., *Democratic* and *Republican*) with the approximated accuracy of 60%.

Here we extend the notion of just detecting adult images from non-adults images into a more formal process by incorporating the notion of sexual intentions behind images. By analyzing the facial expressions, posture, and gesture boosted by scene context and the amount of skin exposure, we try to infer the mood and emotion of the subject, which is equivalent to "persuasive intents" as defined in [11]. In essence our model understands the communicative difference between an almost naked person without any sexual intention (e.g. a bodybuilder or girl in bikini) from a well dressed person with e.g. a seducing facial nudge.

3. Approach

Despite the great progress in computer vision due to deep networks, it is still challenging to infer high-level concepts from raw images or low-level features. Thus, we propose a *hierarchical* method, each level of which captures more abstract concepts than the previous level. We first describe the individual cues that our method combines, then the hierarchical method as a whole. Table 1 presents an overview of our features. We gather annotations for the semantic ones on Amazon's Mechanical Turk platform (see Section 4.1), train models to predict each feature, and at test time use the predicted value rather than ground-truth annotations.

We first extract automatic features (Sec. 3.1) from the images and train multiple, multi-class SVMs with these features to predict mid-level attributes (Sec. 3.2). These posture/gesture, facial expression, scene context and skin exposure attributes in turn are input as features to multiclass SVMs to predict mood and emotion (Sec. 3.3). Then, finally using mood and emotion as features we predict the sexual intent (Sec. 3.4) behind the images.

Feature	Dimensionality
Color histogram	256
SIFT	256
HOG	128
CaffeNet-FC6	4096
CaffeNet-FC7	4096
CaffeNet-FC8	1000
Attributes (posture, gesture, facial expression, scene context, and skin exposure)	17
Mood and emotion	5
Sexual intent	1

Table 1: Dimensionality of our features

3.1. Automatically extracted features

Low-level features are derived from an image using feature detection and extraction techniques, and they capture non-semantic image content such as gradients, edges, etc. Features extracted from convolutional neural networks (CNNs) capture object-part-like templates. We experiment with using each of these features as an image representation from which we learn to predict the attributes.

1. We use a color histogram, as well as the standard Histogram of Oriented Gradient (HOG) [3] and Scale Invariant Feature Transform (SIFT) [15] features.

2. We also use the activations from the fully connected layers of the CaffeNet [9] convolutional neural network, specifically FC6, FC7, and FC8. CaffeNet is pre-trained on 1000 image categories from the ImageNet 2012 visual recognition challenge [17].

3.2. Attributes (Posture, gesture, facial expression, scene context, and skin exposure)

There are many types of visual cues from which we can predict the intent of the subject. Following [11], we extract features which capture the posture, gesture and facial expression of the subject, as well as the scene context of the photograph. We also compute the amount of skin exposure, as this is likely to be a useful cue for our problem. We summarize these features in Table 2, and elaborate on them below.

3.2.1 Posture and gesture

Body cues provide a useful medium for understanding the person's body position, movement, and pose. Body posture can provide significant amount of information about nonverbal communications and emotional signs. Similarly, gesture expresses an idea or meaning via leg, hand, head, or

Posture and Gesture	Facial Expressions	Scene Context	Skin Exposure
Body **P**osture	**L**ooking	Outdoor **S**cene	Fully **C**lothed
Body Position and **M**ovement	**E**yebrow	Outdoor **E**vent	**B**are Bodied
Body Facing Camera	**S**mile	Indoor Scene with **P**rops	**P**rivate Body Parts are Exposed
Face Facing Camera	Eye Li**d**s	Indoor Scene with **F**lat Background	
Head Position	**M**outh		
Spread Eagles Position	**B**iting lips		
Hands Behind Head with **E**lbow			
Position of Hands/**W**rists/Palms			
Gesture with Fingers			

Table 2: Attributes capturing body posture, gesture, facial expression, scene context, and skin exposure.

other body part movements and provides a useful channel to show thoughts, intentions, and feelings through performing physical behaviors.

We characterize the posture cues as the attributes listed as follows: i. Body posture (Straight/Firm $\langle 1 \rangle$, Body arch $\langle 2 \rangle$, Crawling (doggy) $\langle 3 \rangle$, Sitting with folded knees (either stretched or closed)/Frog tie) $\langle 4 \rangle$), ii. Body position and movement (Standing $\langle 1 \rangle$, Sitting $\langle 2 \rangle$, Lying $\langle 3 \rangle$, Walking/Running $\langle 4 \rangle$), iii. Body facing camera (Towards $\langle 1 \rangle$, Away $\langle 0 \rangle$), iv. Face facing camera (Towards $\langle 1 \rangle$, Away $\langle 0 \rangle$), and v. Head position (Straight $\langle 0 \rangle$, Tilted Up $\langle 1 \rangle$, or Tilted Down $\langle 2 \rangle$). We find evidence in our dataset that human subjects in sexually provocative images demonstrate these postures.

The following are some useful gestures that we define for the problem domain: i. Spread eagles position (arms and/or legs stretch) (Not applicable $\langle 0 \rangle$, Hands $\langle 1 \rangle$, Legs $\langle 2 \rangle$, Both hands and legs $\langle 3 \rangle$), ii. Hands behind the head with elbows pointing (Not applicable $\langle 0 \rangle$, Up $\langle 1 \rangle$, Down $\langle 2 \rangle$), iii. Position of hands/wrists/palms (Straight or not applicable $\langle 0 \rangle$, Bent and covering private upper/lower body part (groping/lowering undergarments) $\langle 1 \rangle$, Hands crossed fully (e.g. covering chest, wrists under armpit) $\langle 2 \rangle$), and iv. Gesture with fingers (fingers around face/in mouth/hovering lips/chins) (Yes $\langle 1 \rangle$, No $\langle 0 \rangle$).

3.2.2 Facial expressions

Facial expressions represent motions or positions of the face muscles which are placed beneath the face skin. These non-verbal communications convey social information between people and show the experienced emotion in a specific moment of a situation that could be recorded in an image. The following are the facial expressions which we adopt as signs of sexual innuendos: i. Looking (Away $\langle 1 \rangle$, Straight $\langle 2 \rangle$, Up $\langle 3 \rangle$, Down $\langle 4 \rangle$), ii. Eyebrows (Straight/normal $\langle 1 \rangle$, Frowning $\langle 2 \rangle$, Raised up $\langle 3 \rangle$), iii. Smile (Not smiling $\langle 0 \rangle$, Duchenne $\langle 1 \rangle$, Non-duchenne $\langle 2 \rangle$), iv. Eyelids (Closed $\langle 1 \rangle$, Fully/wide open $\langle 2 \rangle$, Shrunk $\langle 3 \rangle$, Wink $\langle 4 \rangle$), v. Mouth (Open $\langle 1 \rangle$, Closed $\langle 0 \rangle$), and vi. Biting lips or tongue out (Yes $\langle 1 \rangle$, No $\langle 0 \rangle$).

3.2.3 Scene context

As argued in [11], communicative intents can be inferred from the background of an image as it gives out the context and the situation the subject was placed at. We defined this attribute to have one of the following four values: Outdoor scene (natural backdrop of mountain, forest, beach, or any place of seclusion) $\langle 1 \rangle$, Outdoor event (public places like gym, playground, red carpet, swimming pool, etc.) $\langle 2 \rangle$, Indoor scene with props (chairs, couches, curtains, toys, etc.) $\langle 3 \rangle$, and Indoor scene without props and with flat background $\langle 4 \rangle$.

3.2.4 Skin exposure

Although the amount of skin exposed for the subject and position of exposed area with respect to body trunk are often used to classify adult content, it is tricky to differentiate between sexually suggestive content and non-provocative content with the same amount of skin-exposure. We have carefully chosen possible values for this feature as followings: Fully Clothed $\langle 1 \rangle$, Bare bodied (bikini shots or topless) $\langle 2 \rangle$, and Private body parts exposed $\langle 3 \rangle$.

3.3. Moods and emotions

A subject can convey thousands of different types of emotions or feelings that we can attribute to classify the image. To simplify the problem, we choose a series of quantized dimensions of moods and emotions. Such attributes can be one or more of the following: Defensive, protective, or shy $\langle 1 \rangle$, Suggestive, sly $\langle 2 \rangle$, Playful, naughty, or teasing $\langle 3 \rangle$, Happy or relaxed $\langle 4 \rangle$, and Upset, annoyed, angry, or disgusted $\langle 5 \rangle$. These attributes are also shown in Table 3. These moods can essentially capture the emotions and effectively differentiate a sexually provocative image from a non-sexual one.

Defensive or protective or shy
Suggestive or sly (pretension to be shy)
Playful or naughty or teasing
Relaxed or happy
Upset or annoyed or angry or disgusted

Table 3: Mood and emotion of the subjects.

Yes, definitely sexually provocative
Maybe, implicit or hidden sexual intentions
No, casual without any explicit sexual intentions

Table 4: Sexual intent, the global classification task in our framework.

3.4. Sexual intent

This represents the final classification of an image based on the perceived intents. We consider the following three possibilities for whether an image portrays sexual intent or not: Yes $\langle 1 \rangle$, Maybe $\langle 2 \rangle$, No $\langle 3 \rangle$. This final layer of classification is shown in Table 4. In the experiments, we treat intent prediction as a binary classification problem, where a positive classification corresponds to responses of Yes, and a negative classification corresponds to responses of Maybe or No.

3.5. Hierarchical framework

In this section, we lay out the features, defined earlier, in multiple hierarchies to learn sexual intents behind images. Figure 2 represents our proposed hierarchical framework, where one of the many automatically extracted features forms the base of pipeline. There are two intermediate layers of the pipeline, namely our attributes and moods. The top layer is the global sexual intent classification. The character labels used in Figure 2 are obtained from the underlined feature descriptors defined in Tables 2, 3, and 4, respectively.

We use the ground-truth annotations for the above features obtained from Amazon's Mechanical Turk (MTurk) to train our hierarchical framework. Each layer of our hierarchical framework consists of multiple, multi-class classifiers. At test time, we use the predicted values of the lower level features from these classifiers instead of ground-truth values to predict the features at higher level. For all prediction tasks, we use polynomial kernel SVMs with misclassification cost $C = 1$ and polynomial degree 2, 3, and 4, respectively from automatic features to attributes, from attributes to moods, and from moods to sexual intent. Note that cross-validation is used to select the parameters.

Raw image features defined from pixel intensities are used to the learn body posture, gesture, facial expressions, scene context, and skin exposure. The ground-truth (at training time) or predicted (at test time) values for these attributes are used in turn as features for classifiers which predict the moods and emotions of the subject. In turn, the ground-truth or predicted (at training and test time, respectively) moods and emotions values are used to predict the top-level global label, i.e. whether the image portrays sexual intent or not.

4. Experimental Validation

In this section, we first present the new dataset we collected for predicting subtle sexual intent beyond nudity. We then describe how we evaluated our approach, and our results.

4.1. Dataset

We present a new dataset containing 1,146 celebrity images annotated for attributes (pose, expression, skin, and background), moods and emotions, and sexual intents.[2] From *people.com*, 203 Hollywood celebrities are identified based on their popularity. For these celebrities, 1,146 images are collected from *pinterest.com*. Out of these 1,146 celebrities images, there are 892 and 254 images from female and male candidates, respectively. On average there are 5.6 images per person.

The dataset focuses on celebrity images as we observed the presence of sexual cues more for such persons than for average individuals, due to their professions revolving around show-business, modelling, advertisement, professional photo-shoots, and acting. Moreover, the Internet community and general viewer base are more curious and interested in celebrity profiles, which makes it more demanding to characterize such celebrity images for sexual intentions. To attain good classifier performance, we actively tried to maintain balance between imagery with sexual and non-sexual content.

The collected images were annotated on the MTurk platform. Attributes, moods and intents form a set of 19 questions (both yes/no and multiple choice) for each image that each annotator answered. In the process of designing the questionnaire we tried to follow concrete and straightforward descriptions.

To avoid any potential inconsistency in sensitive data, the collected answers from the annotators are further processed for basic data sanity (like incomplete questions). Hits with incomplete annotations were ignored and the questions were re-posted to get further annotations from different annotators. Each image had been annotated by multiple inde-

[2]The size of datasets in [5, 11, 12] are 1,650, 650, and 1,124 images respectively, which is comparable to our dataset size which is 1,146. The only work that uses a larger dataset is the 69,260-image dataset of [7], which belongs to Google's adult content filtering infrastructure.

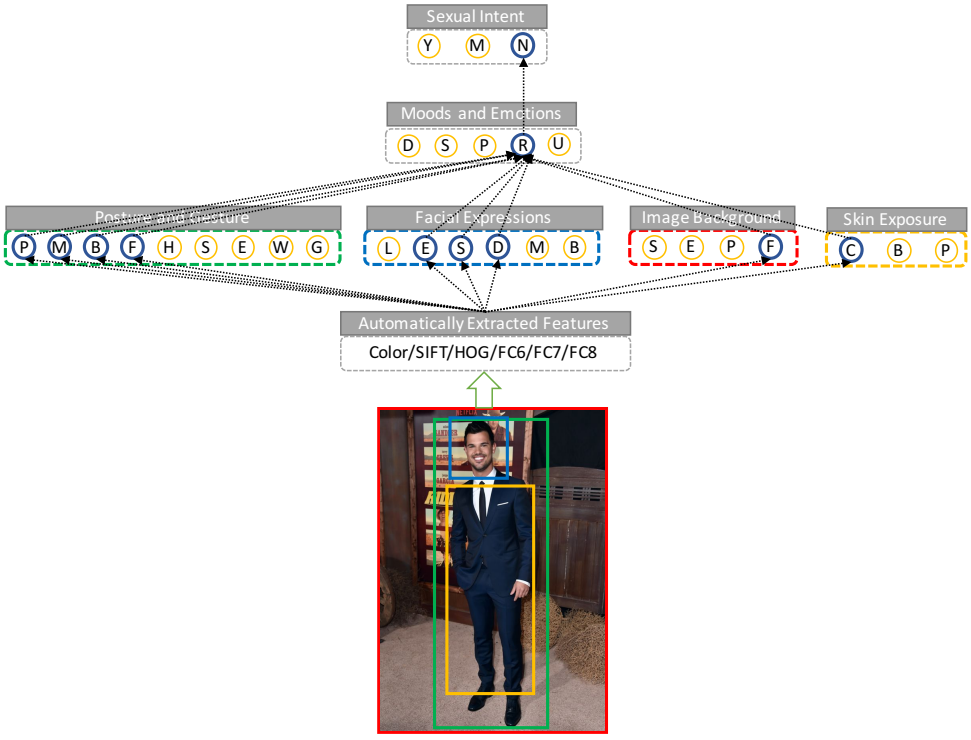


Figure 2: Visual representation of our hierarchical framework. The circled acronyms for features under attributes, moods and sexual intents are derived from highlighted characters in Table 2, 3 and 4, respectively.

pendent annotators as these cues are dependent on the personal beliefs and perceptions of the annotators. We used majority voting of the opinions for our final annotations.

We measured, for each image that was annotated with some label, what fraction of the annotators provided the majority-vote label (i.e. a ratio of 1 indicates complete agreement of annotators over a label). We found that our annotators had consensus of 0.705% on average.

4.2. Baselines

We compare the *Hierarchical Framework*, described above, to the *Direct Model* which serves as the baseline approach for adult content recognition. In contrast to hierarchical model, the direct model is composed of a single level of classification hierarchy i.e. predicting sexual intent directly from *automatically extracted features*. Like the hierarchical pipeline, the direct model uses an SVM with polynomial kernels.

We also compare against the hierarchical framework of [11]. Originally they used the software from [19] to produce their facial expressions, whereas to recreate their model we used [18] to extract the same features from the face bounding box. Note that [18] can detect faces on only 886 images in our dataset, so we only work with these images while collecting results for [11]. Also, [11] used the interface from [4] to create the posture, gesture and scene context raw features, and we followed a similar approach as described by Lazebnik et al. [13] to extract posture and gesture raw features from the rectangular bounding box of the person of interest and scene context from the entire image. Note that we have used tools presented in [18] and [13] instead of [19] and [4] respectively as the latter software are not publicly available. To have a uniform comparison framework, we implemented the concept of Joo et al.'s method in two different approaches: direct and hierarchical. For the direct implementation, we concatenated all available features including facial expressions, gesture and posture, and scene context from the entire image to create a single feature vector to train a classifier to predict the top level sexual intents. For the hierarchical implementation, we followed the hierarchical prototype as presented in their paper closely. Note that the original feature set of [11] had 15 dimensions; but out of those, only 9 features closely match with our feature set and are relevant for our problem scope. We used these features along with the 3 attributes extracted using tools in [18] and [13] to constitute a 12-dimensional feature set for [11]'s hierarchical model. Compared to our hierarchical model, these attributes are not learnt from CaffeNet FC features and thus the results are not directly comparable to the results obtained from our hierarchical model.

Statistics	Joo [11]		Color Hist.		SIFT		HOG		FC6		FC7		FC8	
	Dir.	Hier.	Dir.	Hier.	Dir.	Hier.	Dir.	Hier.	Dir.	Hier.	Dir.	Hier.	Dir.	Hier.
Specificity	0.23 ± 0.08	0.29 ± 0.19	0.30 ± 0.07	0.13 ± 0.04	0.28 ± 0.03	0.14 ± 0.05	0.34 ± 0.07	0.15 ± 0.04	0.43 ± 0.06	0.28 ± 0.06	0.43 ± 0.05	0.28 ± 0.06	0.45 ± 0.06	0.27 ± 0.06
Sensitivity	0.52 ± 0.06	0.47 ± 0.30	0.43 ± 0.09	0.83 ± 0.04	0.47 ± 0.08	0.75 ± 0.06	0.40 ± 0.08	0.70 ± 0.08	0.71 ± 0.07	0.83 ± 0.04	0.66 ± 0.07	0.83 ± 0.06	0.59 ± 0.08	0.84 ± 0.08
Accuracy	0.35 ± 0.06	0.37 ± 0.05	0.35 ± 0.05	0.41 ± 0.04	0.36 ± 0.04	0.39 ± 0.04	0.36 ± 0.05	0.37 ± 0.05	0.54 ± 0.05	0.50 ± 0.05	0.52 ± 0.04	0.51 ± 0.05	0.51 ± 0.03	0.50 ± 0.06
F-measure	0.52 ± 0.06	0.36 ± 0.20	0.35 ± 0.06	0.53 ± 0.04	0.37 ± 0.06	0.50 ± 0.04	0.34 ± 0.07	0.47 ± 0.06	0.56 ± 0.06	0.57 ± 0.05	0.53 ± 0.05	0.58 ± 0.05	0.49 ± 0.05	0.58 ± 0.06

Table 5: Performance of all features when used to predict the intermediate and eventually top layer of our hierarchical model ("Hier.") compared to when used directly to predict intent ("Dir."). We show the mean and standard deviation over all test images.

Model	Specificity	Sensitivity	Accuracy	F-measure
Ground-truth attributes to sexual intent	0.37 ± 0.06	0.83 ± 0.05	0.55 ± 0.04	0.60 ± 0.04
Ground-truth moods to sexual intent	0.41 ± 0.04	0.84 ± 0.05	0.59 ± 0.03	0.62 ± 0.03

Table 6: Performance of the direct models trained from ground-truth annotations of attributes and moods to predict the top-level classification of sexual intent.

4.3. Performance metrics

As we are treating the top layer as a binary problem for evaluation purposes, the true condition is evaluated against predicted condition where True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) identify different correct or incorrect results. Hence, we compute the following metrics: Specificity ($\frac{TN}{TN+FP}$), where lower specificity indicates higher false positive rate; Sensitivity ($\frac{TP}{TP+FN}$), where lower sensitivity indicates higher false negative rate; F-measure ($\frac{2\times Precision\times Recall}{Precision+Recall}$); and Accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$).

We report results using 10-fold cross-validation.

4.4. Predicting global sexual intent

In this section, we evaluate the performance of all automatically extracted features (color histogram, SIFT, HOG, FC6, FC7, and FC8). We also compare the performance of directly using these features to predict intent, against using them to learn the intermediate levels of our hierarchical model and also the state of art hierarchical model described in [11]. The results are shown in Table 5.

Our hierarchical model improves upon the direct model in most cases, except for specificity. The accuracy of our hierarchical model is comparable to that of the corresponding direct model, while our F-measure is much higher than the corresponding direct model. The highest accuracies and F-measure belong to the CaffeNet FC features. For our hierarchical model, these are more than 50% and 57%, respectively. CaffeNet FC7 and FC8 have the highest F-measure of 58% for the hierarchical model, and FC6 has the highest F-measure of 56% for the direct model. Using FC7, we compromise on accuracy by 1% in case of the hierarchical model compared to the direct model; but we gain on F-measure by 5%.

Further, for most metrics, both of our direct and hierarchical model with CaffeNet FC7 outperform the direct and hierarchical models using the attributes from [11]. Moreover, the F-measure for the hierarchical model of [11] shows a very high variance of 20% which signifies it is not a robust approach to identify sexual intent compared to our hierarchical model.

To establish how the ability to predict attributes and moods affects the top-level intent predictions, we also conducted experiments where the *ground-truth* annotations for all attributes, moods and emotions were used as features in the prediction of sexual intent. We show the results in Table 6. One row shows directly predicting top-level intent from ground-truth attributes, and the other show directly predicting intent from ground-truth moods. We see both of the direct models trained from ground-truth annotations of attributes and moods outperform our hierarchical model with CaffeNet FC7 in all metrics, as expected. In our future work, we will investigate ways to complement the data used to predict attributes and moods via external annotation sources or models, so that we can predict these mid-level features more accurately.

To put these results in more perspective: our analysis suggests that the hierarchical model is better than the direct approach because it is more sensitive and and less specific, i.e., it detects more positives. Lower specificity means higher false positive which leads to classify non-provocative images as sexually-provocative images. The specificity of our hierarchical model is much lower than the corresponding baseline for all feature descriptors. In a real world application of this model, the lower specificity imposes a manual refining process for detecting non-sexual contents out of sexually classified contents. If we wish to prioritize the abil-

Attribute	Accuracy
Body **P**osture	0.82 ± 0.03
Body **M**ovement	0.78 ± 0.03
Body Facing Camera	0.79 ± 0.04
Face Facing Camera	0.87 ± 0.03
Head Position	0.82 ± 0.06
Spread Eagles	0.94 ± 0.03
Elbow Pointing	0.94 ± 0.02
Position of **W**rist	0.84 ± 0.03
Gesture with Fingers	0.94 ± 0.02
Looking	0.79 ± 0.06
Eyebrow	0.94 ± 0.01
Smile	0.72 ± 0.03
Eyeli**d**s	0.88 ± 0.03
Mouth	0.56 ± 0.03
Biting lips	0.97 ± 0.01
Scene Context	0.76 ± 0.05
Skin Exposure	0.66 ± 0.05

Table 7: Prediction accuracy for attributes using FC7.

ity to catch any and all sexually provocative images, then *a decrease in specificity is less problematic than a decrease in sensitivity*. The sensitivity of the hierarchical model is *much higher* than the direct baseline.

The strength of our hierarchical model is practical from the perspective of an automated image filtering. Assume we have an automated adult-content filtering infrastructure which protect users' experience with the proposed hierarchical pipeline. Nowadays, whenever users upload pictures in certain applications, there is a latency between upload and display. This is the hold time for review by application administrator. In our hierarchical model whenever we declare a picture as negative, we can confidently display them to users immediately without any hold time or may assign it to less expert reviewers to verify the true credibility. In the case when we declare some pictures as sexually provocative, we demand more attention or in other words usual hold time for review. As users are used to the hold time, so higher false positive rate will not hurt the users experience.

4.5. Recognizing attributes and moods

In this section, we evaluate the usefulness of our hierarchical model beyond the global classification task by presenting the performance of the model in predicting features in the intermediate layers. From Table 5 CaffeNet FC7 has the highest accuracy while predicting the global classes among other automatic features, using our hierarchical model. As previously discussed, we train multiple multi-class SVMs from CaffeNet FC7 feature descriptors, to calculate the attributes (posture, gesture, facial expression, scene context, and skin exposure). Next, we evaluate these predicted results against the ground-truth data and report their accuracy in Table 7. Similarly, we use these attributes to predict moods by training multiple multi-class SVMs. The estimated models are checked against the ground-truth data for moods and emotions and the results are reported in Table 8.

Mood	Accuracy
Defensive	0.98 ± 0.01
Suggestive	0.68 ± 0.06
Playful	0.62 ± 0.03
Relaxed	0.62 ± 0.03
Upset	0.97 ± 0.02

Table 8: Prediction accuracy for moods using FC7.

From Table 7, we see that our model performs better for predicting some attributes (e.g. face facing camera, elbow pointing, eyelids) than others (smile, mouth, skin exposure). The lack of accuracy while predicting some attributes can be compensated by using independent classifiers using datasets for individual attributes. Along the same lines, from Table 8, we observe that prediction accuracy of moods is dependent on the ability of our hierarchical model to predict the associated attributes in the lower layer. Hence, we believe using external sources to predict attributes will also boost the accuracy of moods.

5. Conclusion

Successful realization of the proposed methodology demonstrates that a hierarchical approach to sexual intent prediction can yield new insights into the problem which goes beyond traditional classification based on surface features. Integrating the proposed methodology with mobile apps, social media websites, and media streaming websites will enable automated content classification based on behaviours and intentions of human subjects in such hosted multimedia contents. This will help the content administration to take necessary actions on immediate basis or may prompt the system to seek for intervention of human experts to judge such content for any disciplinary actions for the users. We believe our contribution towards the research of identification of human intentions from images in general will open up new dimensions. For example, similar hierarchical models based on facial expressions, body language, and moods and emotions can lead to exploration of identification of criminal activities from surveillance footage.

References

[1] https://zephoria.com/top-15-valuable-facebook-statistics/, Accessed May 21, 2016.

[2] http://www.reelseo.com/hours-minute-uploaded-youtube/, Accessed May 21, 2016.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.

[4] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2010.

[5] T. Deselaers, L. Pimenidis, and H. Ney. Bag-of-visual-words models for adult image classification and filtering. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 1–4, Dec 2008.

[6] L. Duan, G. Cui, W. Gao, and H. Zhang. Adult image detection method base-on skin color model and support vector machine. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 797–800, 2002.

[7] W. Hu, O. Wu, Z. Chen, Z. Fu, and S. Maybank. Recognition of pornographic web pages by classifying texts and images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(6):1019–1034, 2007.

[8] X. Huang and A. Kovashka. Inferring visual persuasion via body language, setting, and deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2016.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[10] F. Jiao, W. Gao, L. Duan, and G. Cui. Detecting adult image using multiple features. In *Info-tech and Info-net, 2001. Proceedings. ICII 2001-Beijing. 2001 International Conferences on*, volume 3, pages 378–383. IEEE, 2001.

[11] J. Joo, W. Li, F. F. Steen, and S.-C. Zhu. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 216–223. IEEE, 2014.

[12] J. Joo, F. F. Steen, and S.-C. Zhu. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3712–3720, 2015.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178. IEEE, 2006.

[14] J.-S. Lee, Y.-M. Kuo, P.-C. Chung, and E.-L. Chen. Naked image detection based on adaptive and extensible skin color model. *Pattern Recognition*, 40(8):2261–2270, 2007.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[16] H. A. Rowley, Y. Jing, and S. Baluja. Large scale image-based adult-content filtering. In *VISAPP*, pages 290–296. Citeseer, 2006.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[18] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3476–3483, 2013.

[19] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013.

[20] Q.-F. Zheng, W. Zeng, W.-Q. Wang, and W. Gao. Shape-based adult image detection. *International Journal of Image and Graphics*, 6(01):115–124, 2006.