1J4-OS-18a-6in

国内ウェブサービスでの投稿画像検閲における HC/CS 活用事例

A hybrid approach of human computation and machine classification for image moderation

林 佑樹[*1]
Yuki Hayashi

横川 毅[*2]
Tsuyoshi Yokokawa

[*1] 株式会社 IkkyoTechnology
Ikkyo Technology Inc.

[*2] 株式会社 IkkyoTechnology
Ikkyo Technology Inc.

With the rapid expansion of smartphone usage globally, there are more social web services and apps that need to moderate uploaded visual contents by users, "User Generated Visual Contents". We are currently working with multiple online social services that have more than 10 million users, so that we can optimize their visual contents moderation effort. In this report, we are going to describe how we are significantly reducing visual contents moderation cost while maintaining high level of accuracy through combining human-computation with machine-learning.

1. はじめに

インターネットにおける投稿画像の数は年々増加傾向にあるにも関わらず、未だ固定人員による目視検閲が行われている。一方でコンピュータービジョンや深層学習等の技術進歩により画像の自動認識性能が向上してきているものの、100%の精度が求められる用途においては依然として目視による補完が必要である。

そこで我々は、画像の自動認識技術とヒューマンコンピュテーションを組み合わせたワークフローを、機械学習で最適化することで、実用的な検閲精度を確保しつつコストも削減する手法を設計し、検証した。本報告書では、実際に数千万人のユーザーを抱える大手ウェブサービスに対して、提案手法を適用した事例のうち数例を紹介する。

2. 現状分析

ソーシャルメディア上でのやりとりが発端となる事件が社会的な問題となっているが、サービス内での検閲レベルは各運営者の自己責任に委ねられており、ユーザーからの通報に応じて対応する消極的な体制が主流である。特に画像検閲については、国内の主要 BPO 各社が人海戦術による目視確認に依存していることから、ごく一部の不正な投稿のために膨大なコストを掛けて全数検査を行うことが現実的に難しいという声がサービス運営者から上がっている。

しかし、一方では社会的責任を追及されるリスクの高い上場企業や、ブランドイメージをビジネスの根幹とするようなサービスにおいては、採算度外視で検閲に取り組む例も少なくない。国内に限定しても、一つのサービスの画像検閲にかかる費用が毎月数百万円(売上全体の30%以上)に及んでいるケースも存在した。また、サービスをグローバルに展開する上で各国の法律や宗教、文化による厳しい規制への対応が不可欠である他、今後開拓が進むであろう子供向けのサービスを始めるにあたっても避けられない問題として表面化してきている。

また、サービス運営者によっては、検閲対象が漏洩する可能性を嫌い、社内の人員や BPO 先でのセキュアに管理されたワーカーに拘る声も根強く、安価な海外のクラウドソース型検閲サービスの活用にも課題が残る。

連絡先:株式会社 Ikkyo Technology (info@ikkyotech.com)

3. 先行事例

人海戦術によるソリューションと、機械分類によるソリューションに大別出来る。ちなみに本稿執筆時点では、これらを組み合わせた公知のサービスは見つからなかった。

国内においては、固定人員での人海戦術によるソリューションが主流であり、個別の交渉による段階的な従量課金方式が採られている。一方海外では、数百万人のワーカーを抱えたクラウドソーシング型のプラットフォームも存在する。機械分類によるソリューションは一枚あたりの単価が人海戦術に比べて安価であるが精度と柔軟性で劣る。

検閲業務自体をサービスとして一般に提供するのではなく、社内のワークフローを効率化する一環として実施しているケースが存在する。また、動画検閲において特徴フレームを抽出して一覧性の良いサムネイル画像としてワーカーに提示することで検閲を効率化する手法等が提案されている。

4. 要求される検閲事項

サービスによって要求される検閲事項は様々であるが、以下に代表的な例を挙げる。

- 性表現を含む画像
- 暴力表現を含む画像
- 個人情報を含む画像
- 子供の顔を含む画像
- 酒類や薬物を含む画像
- 宗教的タブーを含む画像
- 他者の権利を侵害する画像
- 手書き画像と写真の区別

またこれらの項目に加えて、それぞれ程度の差による細かい基準を設け、コミュニティの活発性を保てるようローカライズ先の地域によって調整される。例えば、宗教上の理由から皮膚の露出が規制されるケースや、個人情報の定義の違いによる国や文化圏毎の違いに加え、匿名掲示板サービスのように表現に極力自由度を持たせることをサービスの売りとしているケース等、柔軟な対応が求められる。

5. 機械による分類

機械による分類で活用出来る基礎技術の例を以下に挙げる。

- 文字認識（OCR）
- 特定物体認識（パターン認識）
- 大規模分類（特徴点抽出・クラスタリング）
- その他、多変量解析手法による分類

6. 人力による分類

6.1 タスクの最適化

タスクのインターフェースを設計する際に、人間の特性を考慮することで作業効率の向上が期待出来る。具体的な例を以下に挙げる。

- 視覚的特性
  - 残像効果の活用
  - 周辺視野の活用
  - 視点移動の最小化
  - 明瞭なカラーリング
- UI（ユーザーインターフェース）
  - クリック数の削減
  - タッチインターフェースの活用
- タスクの規模
  - 作業内容の細粒化

6.2 ワークフローの最適化

タスクをワーカーに割り振る際に、各々のワーカーの能力やタスクへの適正を考慮した上でタスクを順序、粒度、冗長性などを適切に調整することで、作業効率の向上が期待出来る。

6.3 心理的負担の軽減

長時間の検閲作業に従事するワーカーにおける心理的負荷を考慮しなくてはならない。

例えば、対象に卑猥な画像が多いケースにおいては、画像自体に断片化やぼかし等のエフェクトを施し、作業に必要な最低限のクオリティを保ちつつディテールを落とした表示を行うことで心理的負担の軽減が図れる。

また、ワーカーの性別や宗教を考慮して予めタスクの割り当てを制御することで更にリスクを排除することが出来る。

7. 検証結果

7.1 ケース A：アニメーション投稿サービス

- 要求事項
  - 卑猥な画像の除外
  - 個人情報（SNSアカウント・電話番号等）の除外
  - クラウドソーシングの活用
- 適用手法
  - 文字認識（OCR）
  - 特定物体抽出（パターン認識）
  - 大規模分類（特徴点抽出・クラスタリング）
  - ワークフロー最適化
- 結果

  全体で 60%~90%のコストが削減された。不適切なコンテンツが全体で占める割合は約 5％であったものの、アニメーションの中での重複フレームの排除や、偽陽性（false positive）を最小化するようにチューニングした分類器により効率的な枝刈りが出来たことに加え、類似画像のグルーピングと優先度設定が約 3 倍以上の作業効率向上を実現した。

7.2 ケース B：アイコン投稿サービス

- 要求事項
  - 重複画像の排除
  - 写真画像の排除
  - 他者の権利を侵害する画像の排除
  - 社内のワーカーの割り当て
- 適用手法
  - 特定物体抽出（パターン認識）
  - 大規模分類（特徴点抽出・クラスタリング）
  - ワークフロー最適化
- 結果

  まず完全一致及び類似画像検索のアルゴリズムによって、全体の 65％のコンテンツが自動排除出来た。この違反コンテンツの割合の多さはサービス特有のユーザー動向によるところが大きい。しかし、従来の固定人員の記憶力に依存した検閲手法では担当者間の知識共有も出来ない等の理由から見落としは避けられず、悪意を持ったユーザーがその隙を狙った投稿をしていた可能性があるため、本手法の適用によりユーザー体験の向上も期待できる。

  また、ワーカーのタスク生成においても、比較的小さい固定サイズのアイコンは、クラスタリングにより明確な特徴を持ったグループに分かれやすいため、ランダムな検閲と比べて視認性の高いタスクの設計が可能であった。

  これらの施策により、コンテンツ検閲にかかる時間は 90％削減された。

8. 今後の展望

本報告書では、国内の大規模サービスにおける画像検閲のためにヒューマンコンピューテーションを有効活用している実例を示した。また、コンピュータービジョンや機械学習の技術を用いた前処理の有効性も示した。

ヒューマンコンピュテーションの観点では、今後機械による認識性能の向上に伴って、人間にしか出来ないことの見極めがより重要になってくるものと考えられる。

また、コンテンツ監視の観点では、今後より拡大するニーズに対応するために、検閲フレームワークの汎用化と、サービスを横断したデータベースの構築が必要であると考えられる。それにより、サービスによって異なる検閲対象の画像特性や、要求される検閲事項の違いにも迅速に対応出来るようになる他、識別器の流用可能性が期待できる。また、保護すべきコンテンツの情報を一度登録すれば自動的に複数のサービスで検閲対象となるような共通のプラットフォームが出来れば、業界全体の健全化につながるだろう。