DEIM Forum 2021 D14-4

画像におけるゲームによる多様・信頼的・効率的な感情アノテーション

左　幸坤† 　李　吉屹† 　茅　暁陽†

† 山梨大学コンピュータ理工学科　〒 400–8511 山梨県甲府市武田 4 丁目 3-11
E-mail: †luckykunzuo@gmail.com, ††{jyli,mao}@yamanashi.ac.jp

あらまし　この原稿は，我々の研究チームが ACM Multimedia 2020 に発表したフル論文 (口頭発表)「AffectI: A Game for Diverse, Reliable, and Efficient Affective Image Annotation」[10] の要約と紹介である．近年，ディープラーニング技術により，信頼性の高い大規模な学習データセットを構築するために，感情的な画像アノテーション技術の需要が高まっている．提案手法では，Game With a Purpose (GWAP) の概念に基づいて，多様で信頼性の高い感情ラベルを効率的に収集するための新しい感情画像アノテーション技術 AffectI を提案する．AffectI は，多様で信頼性の高いラベルを収集するために，3 つの新しいメカニズムを提案している：(1) 全ての感情語を公平的に評価する選択メカニズム；(2) ラベルを効率的に収集するために部分的な感情語のペアワイズ比較を統合して，感情分布を推定する推定メカニズム；(3) 現在のプレイヤーと相手や過去のプレイヤーとの比較を表示して，プレイヤーの興味を喚起するインセンティブメカニズム．実験結果について，AffectI は，より多様で信頼性の高いラベル収集が可能であるという点で，既存手法よりも優れていることが示された．また，プレイヤーの不満を軽減するために GWAP を利用するメリットは，主観的な評価を通じて確認された．

キーワード　マルチメディア、感情画像アノテーション、クラウドソーシング、Game with a Purpose

1 はじめに

感性的な画像アノテーションは，画像検索や感性的な画像コンテンツ解析など，幅広い分野に応用が可能である．エンドツーエンドで特徴を学習できるディープラーニング技術は，感性的な画像のコンテンツ解析において大きな注目を集めており，信頼性の高い大規模な学習データセットを構築するために必要な感性画像アノテーション技術の需要が高まっている．

既存の大規模な感情画像データセットは，主にソーシャルネットワークからの画像を使用して，自然言語処理と手動ラベル付けを組み合わせて構築されている [1,7–9]． 例えば，FlickrCC データセット [1] は，3,000 の形容詞と名詞のペア（ANP）を持つ Flickr クリエイティブコモン（CC）画像を取得することによって構築される． MVSO データセット [5] は，FlickerCC データセットの多言語バージョンである．

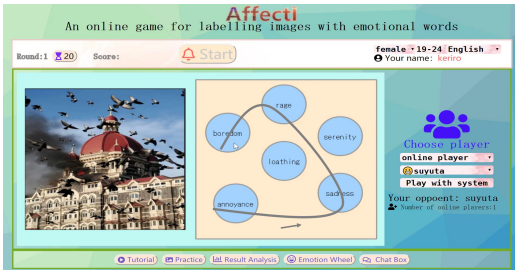
ソーシャルネットワーク上の画像にアノテーションされたテキスト情報を活用することは，大規模なデータセットを収集するのに効率的ですが，この方法で収集されたデータは，説明に含まれる感情的な言葉は，画像が伝えようとしている感情とはあまり関係がないかもしれません．このようにして収集されたデータは，通常非常にノイズが多いである．収集されたラベルは，高頻度の感情的なラベルに偏っているため，多様性と品質が不足している．

提案方法では，Game With a Purpose（GWAP）の概念に基づいて画像の感情ラベルを収集するため，新しい感情画像アノテーション手法 AffectI を提案する．提案するシステムは主に以下の 3 つの構成要素から構成されている．選択メカニズム（図 1.(a)）は，全ての感情語が公正に評価されることを保証し，プレーヤーが効率的に感情語とその程度を選択および判断することを支援する．推定メカニズム（図 1.(b)）は，複数のプレーヤーによって提供された部分的なランクリストから全てのラベルの感情度を推定する． インセンティブメカニズム（図 1.(c) および (d)）は，現在のプレーヤーとその相手，または過去の全てのプレーヤーとの比較を示している．

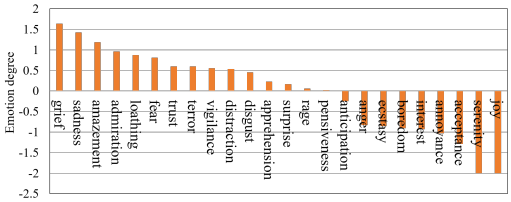
このシステムは，画像の多様なおよび信頼的な感情ラベルを収集する効率的に重点を置いている． 第一に，既存の研究 [5] のように頻度の高い通常の感情語のみを使用するのではなく，多様な感情度を持つ感情ラベルを収集することができる． 第二に，収集した多様な感情ラベルは信頼性が高く，感情分類法（Plutchik ’s Wheel of Emotions [6]）における全感情語の感情分布が提供されており，感情画像のコンテンツ分析に信頼性の高いシステムとなっている．第三に，プレイヤーは，全ての感情語において，小さな部分集合の感情度を部分的に判定するだけでよいので，プレイヤーの作業負担を軽減することができる．AffectI を使用すると，プレーヤーは直感的で楽しく画像にラベルを付けることができる．したがって，システムは効率的である．

2 提案方法

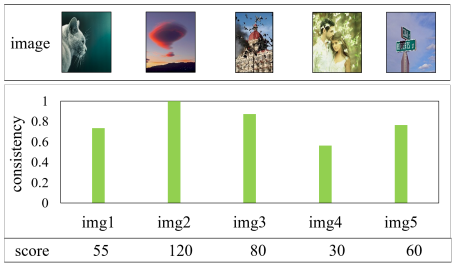
GWAP の概念に基づいて、画像から感情ラベルを収集する感情画像アノテーションのためのオンラインゲームを提案する．AffectI の新規性は，感情の主観的および相対的な知覚の問題に対処するために設計された 3 つの主要なメカニズムを提案している． 1 つ目は選択メカニズムである．これは，全ての感情語が公正に評価され、効率的に単語の感情度を選択して判断することができるようになる．2 つ目は，推定メカニズムである．つまり，Bradley-Terry モデル [2] を適用して，複数のプレーヤーによって提供された部分的なランクリストから全ての単語



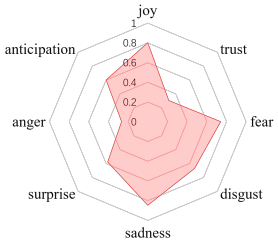
(a). 選択メカニズムのインターフェース：候補となる感情語を公平に選択して表示し，プレイヤーは線をドラッグすることで感情語を感情の度合いの高い順に評価する．



(b). 推定メカニズム：複数のプレーヤから提供された部分ランクリストに基づいて，24 の感情語の感情度を推定する．



(c). インセンティブメカニズム 1：1 回のゲームラウンドでの各画像上の相手とのゲームポイントと一貫性の程度．



(d). インセンティブ・メカニズム 2：8 つの主感情の全ての過去のプレーヤーとの一貫性の程度．

図 1: AffectI の概要

の感情度を推定する．3 つ目はインセンティブメカニズムである．これは，現在のプレーヤーとその相手，または過去の全てのプレーヤーとの比較を示し，プレーヤーに高品質のラベルを提供するように促すことができる．図 1 は，提案された方法の概要を示している．

2.1 選択メカニズム

画像にラベルを付けるために自由な単語の選択が許可されている場合，プレーヤーが同じ単語を選択する可能性はほとんどありません． このように，プレイヤーに感情の単語の候補から選択するように依頼することで，ラベルを収集する．AffectI は，Plutchik's Wheelof Emotions [6] の 24 の感情語を使用している．これは，プレーヤーが感情のニュアンスとそれらの間のコントラストを理解するのに役立つように設計されている． 怒り (anger)，嫌悪感 (disgust)，恐れ (fear)，悲しみ (sadness)，期待 (anticipation)，喜び (joy)，驚き (surprise)，信頼 (trust) の 8 つの主感情があり，各主感情には 3 つの強さのレベル（激しい (intense) から穏やかな (mild) まで）がある．

画像の各単語の感情度を得るためには，プレイヤーに一般的な評価尺度（Semantic Differential や Likert Scales など）を用いて各単語を評価させることが有効である． しかし，プレーヤーが絶対的な度合いを割り振るのは容易ではない．別の方法としては、ペアワイズ比較を使用することで、プレイヤーが判断しやすくなる．ただし，24 語のペアワイズ比較の総数はかなりに多いである．そこで，プレーヤの比較回数を減らすために，感情語のサブセットをプレーヤに比較させ，画像との感情の一致度合に基づいて部分的なランキングリストを与える仕組みを採用する． 図 1.（a）はプレイヤーの選択インターフェースの例を示している． 複数のプレーヤーが同じ画像にラベルを付けた後，感情の程度を推定することにより，これらの部分的なランクリストを 1 つの全体ランクリストに統合する． 推定メカニズムは次のセクションで紹介される．

一方で，画像のラベルとなる確率の高い単語を提示することで，プレイヤーがより質の高い単語を選択しやすくする（開拓 exploitation）；一方で，頻度が低く示されている単語については，感情度をより正確に推定するために，さらに評価する必要がある（探索 exploration）．したがって，我々の選択メカニズムは、探索と開拓のトレードオフを行うハイブリッド戦略である．プレイヤーに提示された 6 つの単語の中から，ある閾値 θ よりも高い確率でラベル付けされる 2 つの単語をランダムに選択する．また，プレイヤーに表示される頻度が最も低い 4 つの単語を選択する．提案された選択メカニズムの決定的な利点の一つは，多様な感情をアノテーションすることをプレイヤーに促し，主感情の中に多様な感情度を持つ単語を提供することである．

2.2 推定メカニズム

プレイヤーに提示される画像は，画像データセットからランダムに選択される．同一の画像を，異なる 6 つの感情語のセットで複数回用いて評価する．上述したように，各プレーヤーは，感情語のサブセットを評価するだけである．したがって，これらの部分的なランクリストを統合して，24 個の感情語の全てについて感情度を推定する方法が必要である．そこで，部分的なペアワイズ比較ラベルから全オブジェクトの順位スコア（全単語の感情度）を推定する方法として，BT（Bradley-Terry）モデルを採用している．BT には CrowdBT [3] のような亜種が

提案されている．これらのバリエーションでは，複数のプレイヤーが同一オブジェクト上で選択する選択肢の矛盾をモデル化する際の前提条件が異なる．CrowdBT では，ラベルが不一致の場合，プレイヤーの能力によってミスをするプレイヤーがいると仮定している．我々の研究では，複数のプレイヤーのラベルの不一致は個人の認識によるものであると仮定している．そこで，この仮定に沿って BT モデルを適応させる．図 1.(b) は，複数のプレイヤーからの部分ランクリストをもとに，画像に対する 24 の感情の程度を推定した例を可視化したものである．

2.3 インセンティブメカニズム

AffectI では，2 種類の異なるインセンティブを提案する．1 つは，現在のプレイヤーと相手を比較することで，プレイヤーに質の高いラベルを提供することを促すもので，これを Opponent-based Incentive (OI) と名付けられている．もう一つは，現在のプレイヤーと過去のプレイヤーを比較したもので,「過去のプレイヤーに基づくインセンティブ (Past Players-based Incentive: PPI)」と名付けられている．実証結果からは，このインセンティブを適用した後，より多くのプレイヤーが大多数のプレイヤーとは異なる多様なラベルを提供する傾向があることがわかりました．

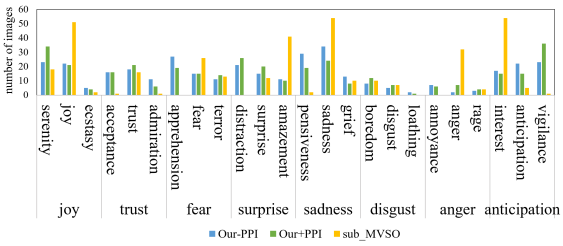
3 評 価 実 験

3.1 実 験 設 定

実験のために，MVSO データセット [5] から 60 枚の画像を選択しました．これは，形容詞・名詞ペア（ANP）を用いて構築した FlickerCC データセット [1] の多言語バージョンである．「美しい花」や「悲しい目」などの ANP は，Plutchik ’s wheel から感情語を用いて検索し，頻度と多様性を考慮して選択した．MVSO には 4,000 を超える ANP があり，各 ANP は複数の画像に関連付けられている． 我々の実験では，8 つの主感情カテゴリのそれぞれに対して，Web ビューの高い 7〜9 枚の画像が選択しました．各 ANP には 24 の感情のスコアがあり，1 つの画像を複数の ANP に関連付けることができる． 感情スコアとして，画像に関連付けられた全ての ANP の平均的な感情スコアを用いた．このようにして，60 枚の画像のそれぞれについて 24 個の感情の離散分布を得た．

ラベルを収集する際のシステムの効率性の問題については，我々の選択・推定メカニズムは自然に感情語の小さなサブセットを迅速に評価し，部分的なランクリストを提供することを可能にしている．これは，自然にプレイヤーの時間コストを削減することができ，したがって，効率的である．

我々のシステムを用いて実験を行いました． 参加者の年齢は 22 歳から 35 歳まで．感情ラベルの収集は，過去のプレーヤーベースのインセンティブ（PPI）のメカニズムを利用した場合と利用した場合に分けて行いました．PPI なしで行われるゲームについては，この設定を「Our-PPI」と名付ける．163 のプレーヤー ID がある．60 の画像には合計 1,892 回のラベルが付けられる．全ての $1,892 \times 6 = 11,352$ 感情語の中から、4,479 個の感情語が選択され，画像にアノテーションを付けた．PPI を使用して行われたゲームについては，この設定を「Our+PPI」と名付ける．プレイヤー ID は 67 個である．60 枚の画像には合計 710 回の感情語が付けられている．全ての $710 \times 6 = 4,260$ 感情語の中から，1,546 個の感情語が選択され，画像にアノテーションを付けた．MVSO データセットのサブセットのアノテーションをベースラインアノテーションとして使用している．



単語頻度 (上位 6 個)

図 2: アノテーションされた感情語の頻度による多様性評価.

3.2 Q1. 多様性

提案したシステムと MVSO で収集した感情ラベルの多様性を比較する．まず，60 枚の画像の上位 k 単語にアノテーションされた感情語の頻度を各システムごとに計算することで，多様性を評価する．これは，24 の感情語すべてが，データセットに含まれる全ての画像のさまざまな程度の感情を表すために適切に使用されているかどうかを評価することである．図 2 は，上位 6 の場合の結果を示している．主感情に基づいて感情の単語をグループしている．

また、MVSO では、上位 6 のアノテーションでは、60 枚中 50 枚の画像に ”喜び ”という感情語がほぼ付与されていることが、図 1 に示されています。つまり、MVSO はコーパスの中で頻度の高い感情語を主に使用しており、多様性に欠けていることがわかります。

実験結果から，sub_MVSO では，主に各主感情の代表的な単語のみを使用しており，激しいと穏やかな感情の単語はほとんど使用していないことがわかりました． 例えば,「喜び」(joy) という主感情については,「喜び」という単語の方が，激しい感情の単語「歓喜的な」(ecstasy) や穏やかな感情の単語「冷静的な」(serenity) よりもはるかに高い頻度である．また，sub_MVSO は，上位 6 のラベルでは，60 枚中 50 枚の画像に「喜び」の感情語をほぼ割り当てていることが，図 2 に示している． つまり，MVSO はコーパスの中で頻度の高い感情語を主に使用しており，多様性に欠けていることを示している．

対照的に,「Our-PPI」と「Our + PPI」では，主感情における感情語の頻度は相対的に一様である． 激しいと穏やかな感情の単語は，画像にラベルが付けられる可能性が高くなる．これは，我々のシステムによって収集された感情ラベルが MVSO よりも多様性が高いことを示している．

第二に，感情語の推定感情度の多様性を評価する．感情語の推定感情度の多様性を，エントロピーに基づく測定を用いて定量的に評価する．画像 a_k に対して，全ての感情語の感情度の

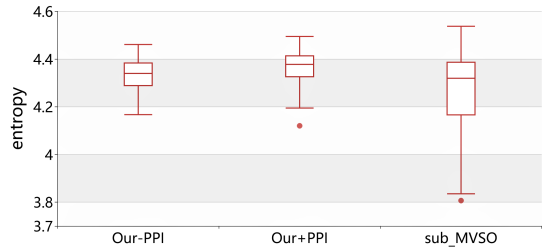


図 3: 24 の感情ラベルに対する各方法のエントロピーの分布.

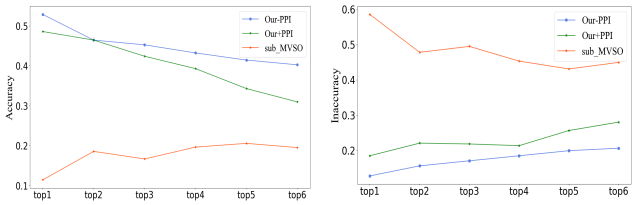


図 4: 上位 k ($k \leqq 6$) の感情語の平均精度と不正確さについて，我々のシステムと sub_MVSO による評価を行った.

エントロピーは， $h_k = -\sum_{i=1}^{24} s_i log s_i$ で計算される．エントロピー値が高い場合は，感情情報がより分散しており，感情ラベルが多様化していることを意味する．図 3 は，各システムの全画像に対するエントロピーの分布を示している．Our-PPI と Our+PPI の両方が sub_MVSO よりも高いエントロピーを持っていることが分かりました．各システムのペアの結果についての t 検定では，$p < 0.05$．統計的に有意な結果が得られた．提案システムが収集した感情ラベルの方が，感情度の多様性が高いことが分かりました.

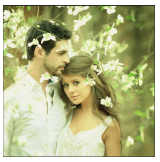
3.3 Q2. 信頼性

感情の程度が異なる多様なラベルを収集できることを検証した．これらの多様なラベルの信頼性を検証する必要がある．データセット全体の画像のアノテーションを評価することは困難であるため，ラベルの信頼性を人が評価するために，提案システムのアノテーションと sub_MVSO のアノテーションとの差が最も大きい 10 枚の画像を抽出した．この 10 枚の画像を 3 つのグループに分け，各グループは AffectI ゲームに参加していない 7 名の評価者によって匿名で評価された．評価者には，本システムと sub_MVSO で得られた感情度に基づいて，上位 6 つの感情語について，不正確性，中立性，正確性の判定を行ってもらう.

図 4 は，上位 k ($k \leqq 6$) の感情ラベルに含まれる感情語の平均精度と不正確さを示したものである．上位 k ($k \leqq 6$) の感情ラベルに含まれる感情語については，我々のシステムの精度は常に高く，不正確さは MVSO よりも常に低い．MVSO では，コーパス中の高頻度の感情語を主に使用しているため，品質が低いだけでなく，信頼性にも問題がある．提案システムが収集した感情ラベルは，画像の感情情報をより正確に記述することができる.

図 5 に示すように，2 つの詳細な例を示している．図 5.(a) では，画像の感情情報には，提案システムでアノテーションされている「期待」(anticipation) が含まれている．逆に，MVSO でアノテーションされている「悲しみ」(sadness) や「驚き」(amazement) などの感情語は，画像の中では目立たないようになっている．図 5.(a) では，提案システムが収集した感情ラベル「心配」(apprehension) のような感情ラベルが，この画像の感情情報をより良く表現することができる．(b) の例では，MVSO が収集した感情ラベルである「狂喜」(ecstasy) や「喜び」(joy) は画像には反映されていません.

さらに，図 5 の例では，MVSO はコーパスに含まれる頻度の高い「喜び」(joy) や「悲しみ」(sadness) などの通常の感情語を用いているのに対し，提案システムは多様な感情語を用いて画像にアノテーションを行うことが可能である．これにより，提案システムが多様で信頼性の高い感情ラベルを収集できることを検証した.



Our-PPI	In.	N.	Acc.
anticipation	0/7	1/7	6/7
interest	0/7	4/7	3/7
joy	0/7	2/7	5/7
sub_MVSO	**In.**	**N.**	**Acc.**
sadness	6/7	1/7	0/7
amazement	4/7	2/7	1/7
joy	0/7	2/7	5/7

(a). 例 1

Our-PPI	In.	N.	Acc.
apprehension	1/7	1/7	5/7
pensiveness	1/7	3/7	3/7
grief	2/7	4/7	1/7
sub_MVSO	**In.**	**N.**	**Acc.**
ecstasy	6/7	1/7	0/7
boredom	3/7	3/7	1/7
joy	6/7	1/7	0/7

(b). 例 2

図 5: 感情的な画像アノテーションの例. In.: 不正確さ; N.: 中立性; Acc.: 正確さ.

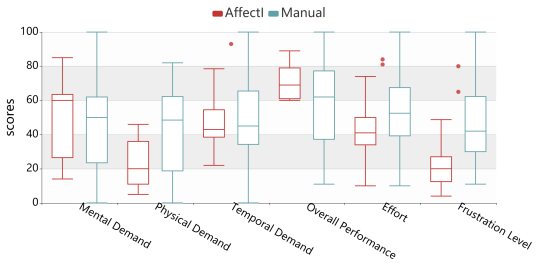


図 6: NASA-TLX 評価結果

3.4 Q3. ユーザーエクスペリエンス

ユーザー体験を評価するために，比較のために 28 名の参加者を対象に手動ラベリング実験を行った．60 枚の画像を 4 つのグループに分け，各参加者は 1 つのグループ（15 枚）の画像に手動でラベリングを行ってもらいました．15 枚の画像のそれぞれについて，24 個の感情語の中から一致する語を選択してもらいました．実験終了後，AffectI のプレイヤー 15 名と手動評価実験の参加者 28 名には，NASA Task Load Index (NASA-TLX [4]) のフォームに記入してもらった．その結果を図に示す．その結果，物理的要求，パフォーマンス，努力，フ

ラストレーションレベルについて，AffectI は手動評価よりも優れていることが分かりました．特にイライラ度の改善が顕著であり，ゲームならではの優位性を示している．

4 ま と め

本原稿では，多様で信頼性の高い感情ラベルを効率的に収集するための新しい感情画像アノテーションシステム，我々発表した論文とシステム AffectI [10] を紹介した． 今後の研究では，データセットを拡張し，画像に対するプレーヤーのパーソナライズされた感情度を推定する．

文 献

[1] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang, *Large-scale visual sentiment ontology and detectors using adjective noun pairs*, Proceedings of the 21st acm international conference on multimedia, 2013, pp. 223–232.

[2] Ralph Allan Bradley and Milton E. Terry, *Rank analysis of incomplete block designs: I. the method of paired comparisons*, Biometrika **39** (1952), no. 3/4, 324–345.

[3] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz, *Pairwise ranking aggregation in a crowdsourced setting*, Proceedings of the sixth acm international conference on web search and data mining, 2013, pp. 193–202.

[4] Lacey Colligan, Henry WW Potts, Chelsea T Finn, and Robert A Sinkin, *Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record*, International journal of medical informatics **84** (2015), no. 7, 469–476.

[5] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang, *Visual affect around the world: A large-scale multilingual visual sentiment ontology*, Proceedings of the 23rd acm international conference on multimedia, 2015, pp. 159–168.

[6] Robert Plutchik, *A general psychoevolutionary theory of emotion*, Theories of emotion, 1980, pp. 3–33.

[7] Yang Yang, Jia Jia, Shumei Zhang, Boya Wu, Qicong Chen, Juanzi Li, Chunxiao Xing, and Jie Tang, *How do your friends on social media disclose your emotions?*, Twenty-eighth aaai conference on artificial intelligence, 2014.

[8] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, *Building a large scale dataset for image emotion recognition: The fine print and the benchmark*, Thirtieth aaai conference on artificial intelligence, 2016.

[9] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, and Tat-Seng Chua, *Predicting personalized emotion perceptions of social images*, Proceedings of the 24th acm international conference on multimedia, 2016, pp. 1385–1394.

[10] Xingkun Zuo, Jiyi Li, Qili Zhou, Jianjun Li, and Xiaoyang Mao, *Affecti: A game for diverse, reliable, and efficient affective image annotation*, Proceedings of the 28th acm international conference on multimedia, 2020, pp. 529–537.