Melanoma Feature Representation between Human Experts and Neural Networks

Murray S. Bennett and Joseph W. Houpt
Department of Psychology
The University of Texas at San Antonio

Abstract

Melanoma is a deadly skin cancer, and early detection is critical for improving survival rates. Dermatologists typically rely on a visual scan to diagnose melanoma by assessing the primary perceptual characteristics of a skin lesion. The common ABCDE heuristic, for example, suggests observers check a lesion for shape (A)symmetry, (B)order irregularity, number of unique (C)olours, and (E)volution over time. Whilst this heuristic provides a practical guide, it is a limited approach. Firstly, all lesions vary and often contain only a subset of these features. Secondly, a combination of abnormal features can lead to a diagnosis, making the diagnostic process complicated and error-prone. Advanced computer vision algorithms (CVA) have emerged as a powerful approach to melanoma identification. CVAs can evaluate lesion features to generate highly accurate and objective assessments. However, despite CVA advancements, they can only be used in conjunction with an expert assessment. Thus, the perceptual expertise of dermatologists remains a critical component in the accurate and timely detection of melanoma. Our project aims to improve the early detection of melanoma by investigating the perceptual judgements of skin lesion colour and shape made by humans and comparing them with the feature representations generated by computer vision algorithms. We recruited non-expert participants online to complete a two-alternative forced-choice task using skin lesion images from the ISIC archive. Participants were instructed to choose the image that exhibited a greater frequency of unique colours in one condition and greater border regularity in another among the two images presented in a trial. We analysed the data using the Bradley-Terry-Luce (BTL) model to estimate each lesion image's relative "strengths" along these perceptual dimensions. We then compared these estimates to computer vision assessments of the same perceptual features. We discuss the methodological approach, preliminary results, and future directions.

Keywords: computer vision, visual perception, melanoma identification, two-alternative forced choice, Bradley Terry Luce model

Research Question / Aim

What we did

What we found

What it means

Take-Home message

Future direction

Complete data collection on the current dimension, and expand to other valued dimensions: colour, size

More advanced computer vision assessment for the evaluation of features and for the segmentation of lesions.

This currently represents a descriptive examination of a single perceptual dimension. Our ambition is to begin an evaluation of features when combined. For example, we could readily expect the perception of border regularity to be affected by lesion size. Whilst we don't have the ground truth for lesion size (e.g., two masks may be the same size because the photographer zooms in, when in reality the two lesions are quite different), we can still provide a proof of concept here by relating the mask size to the BTL scaling.

Actual relationship to melanoma identification – currently only looking at perception of real-world stimuli.

How does a GRT representation actually map to human perception?

Malignant melanoma affects this many people each year, and rates are increasing relative to the population. Early detection of melanoma improves survival rates by a huge amount.

The ability to detect melanoma and differentiate from benign lesions is critical.

General heuristics used by experts, and suggested to the general public, include the ABCDE rule or the ugly duckling rule (if it looks suspicious, it probably is).

However, these are only guides, and no sure-fire method for detecting melanoma, particularly during the critical, but difficult-to-detect early stages of melanoma growth, exists.

Dermatology experience is a clear factor in diagnostic performance.

Machine learning and image processing advances also represent practical and powerful methods for identification of images or collaborative suggestions in conjunction with trained dermatologists.

Whilst the power of machine learning algorithms, image processing, and the list of identifiable features derived by algorithms for machine learning inputs continues to grow, the importance of trained specialists has

Melanoma Identification

Artificial Intelligence

We focus on the ABCD guidelines here as this heuristic lends itself to the approach taught to and used by human observers.

Human Performance

The Current Research

General Recognition Theory

Determination of Appropriate Stimuli

Experiment 1

Image processing for baseline measures were conducted and tested against the ISIC-2017 challenge data, where stimulus performance on image segmentation was the goal. $\approx$ 2000 images with expert-consensus masks were provided for the challenge. We use this dataset as a benchmark for our own image segmentation algorithm, which can then be applied to other images, if need be. I think I might need to because the 2017 data only has melanoma classifications, and ground-truth representations of the melanoma segments. On the other hand, the ISIC database contains size measurements, that can be used to help with the D metric.

Analytical Approach to Feature Identification

Image processing and contour extraction to evaluate shape characteristics, such as border irregularity and asymmetry. There are a number of shape features that can be analysed or assessed in relation to the shape of the melanoma.

Neural Net Approach to Feature Identification

Some neural networks have been trained on/for the ABCD rule. We can use neural networks for image segmentation, the focus of the ISIC-2017 part 1 challenge. We can then take the segmentations and apply the same shape evaluations.

Human Ratings

We can also take human ratings of stimulus features to determine feature scores for each stimulus. This approach to feature rating is subjective, particularly when human-raters are not experts within the field. The issue is compounded by the nature of the stimuli features we aim to rate. That is, perceptual concepts such as symmetry, border regularity, and color spaces are unclear, ambiguous, and other times entirely indiscernible. However, statistical machinery exists that allows the features of a stimulus to be scored on these perceptual dimensions relative to other stimuli. The Bradley Terry Luce model (BTL) estimates the probability that one item is greater than another, for all items in the set. Importantly, not all items of the set are required to be compared to determine a ranking. A novel application, and entertaining example application of the BTL is in the ranking of sporting teams in leagues where not all teams are able to compete against the other.

The model requires an outcome score between items i and j. In the current study, for example, a participant is presented with two skin lesion images and is asked to select the image that is more symmetrical of the two. The selected stimulus is recorded as the "winner" of the match-up between these items.

[[[You could describe some features of the dummy data – e.g., incomplete matchups, asymmetry in matchup scores e.g., A>B, B>A]]]

The BTL estimates the probability that $i > j$, denoted as $P(i > j)$, where $P(i, j)$ is estimated as:

$$P(i, j) = \frac{p_i}{p_i + p_j} \tag{1}$$

The below equation generates the estimate p_i for each stimulus i.

$$p_i = \frac{W_i}{\sum_{j \neq i} \frac{w_{ij} + w_{ji}}{p_i + p_j}} \tag{2}$$

BTL gives us relative image scores. That is, the difference of 0.1 on the scale is consistent regardless of where you are on the board. This is hugely helpful as you can select images that are a relative distance from each other, rather than a similar "score-space" (e.g., 4.3 -> 4.4 is the same as 8.8 -> 8.9).

Methods

Participants

10 million participants ($M_{age} = 35, \sigma_{age} = 10$) were recruited online via MTurk, with participants randomly allocated to one of the four conditions upon entry to the experiment. The final data analysis was conducted on x_1, x_2, x_3, and x_4 participants in the irregular border, consistent border, colourful, and uniform colour conditions. This research was approved by the Human Research Ethics Committee at the University of Texas at San Antonio.

Design

We recorded the competitors in each trial, the winner and loser, and the response time for the decision. The experiment was divided into 4 conditions by isolating two dimensions of interest: border regularity and colour, then accessing each dimension from opposite ends of the feature scale. For example, colour was identified as a meaningful dimension, so participants were presented image pairs and in one condition, asked which was the *colourful* whereas in another condition were asked which lesion had the more *uniform* colour. Similarly, participants were asked to assess border irregularity in one condition and border consistency in another.

Table 1

Experiment Conditions

	Low	High
Border	Consistent	Irregular
Colour	Uniform	Colourful

Results

Discussion

Acknowledgements

Availability of Materials

References