Codebook

Genealogies of Data Content Analysis

□ = structured content analysis
✶ = unstructured thematic analysis

Background

This is the codebook for the Genealogies of Data Content Analysis. This codebook is being used and developed to analyze a ~25% sample of 490 datasets found which are actively used in computer vision research. This work most closely pairs with research question 1 in the genealogies project: *How do dataset developers in CV and NLP research, describe and motivate the decisions that go into their creation?*

Documents

Categories in this section relate to high-level descriptions of both the coding process and the documents being coded.

Documents Analyzed ✶

The documents analyzed for each database. The main purpose of this column is to make it easy to track and find the documents being analyzed, particularly when cross-checking and quality check each other's coding. Any and all links to the following should be included:

- Paper PDFs
- Websites
- Contracts

Coder □

Who coded the specific database.

Sample Strategy □

How the database was sampled. There are 4 options:

- Over 4000 citations = 14
- Body random sample = 22

- Face random sample = 42
- Non-corporeal (non-corp) random sample = 36

Task

Task Found □

The task we originally found to be associated with the database in the CV Corpus. This would be what a research paper used the database to do, but not necessarily what it was intended for. For those labeled N/A, these were found by snowball sampling or were imported from Scheuerman et al. 2020.

Task(s) Intended □

What task(s) authors intended the database to be used for which are enumerated in the paper(s)? Separate tasks with semicolons (;) for ease of parsing.

Domain

Domain(s) Intended □

What real world uses the authors are imagining this database would be used for? Separate tasks with semicolons (;) for ease of parsing.

Domain(s) Intended Thematic

How the authors describe the intended domain(s) in the documents? Paste quotes from documentation into this section.

Contribution

Categories in this section describe the intended contribution of the original research paper.

Is the main contribution the database? □

Is the database presented as the main contribution of the paper? Criteria: The database is presented as a main contribution early and no novel methods or models are introduced. For example, a database is introduced and only evaluated with existing baselines, it may be the main contribution.

This is opposed to situations in which the database is constructed only to test a new method or model.

If the database is not the main contribution, what is the main explicit contribution of the paper?

If the database is not the main contribution, what is? Is a benchmark, a model, or something else the main contribution? We want to know how the authors frame the paper and they see it as the biggest contribution. This is most likely established in the abstract and/or the first section of the paper. But often this is not clear cut, we may make an implicit judgement.

Categories can include:

- N/A:
    - If the main contribution is the database.

- A method:
    - Example: INRIA Pedestrian: We study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks are all important for good results. The new approach gives near-perfect separation on the original MIT pedestrian database, so we introduce a more challenging dataset containing over 1800 annotated human images with a large range of pose variations and backgrounds.

- Empirical Study
    - Example: Forensic Facial Examiner Study Data Set

Is the only thing written about in the paper the database (including descriptive statistics)?

Options: Yes or No.

If the paper contains a model, evaluation on baselines, algorithm, etc., write No.

Is the database described as a benchmark in the original publication?

Whether the authors claim the database is a benchmark. Options: Yes or No.

Is the database included as part of a challenge?

Whether the database was included as part of a challenge (e.g. ImageNet).

Options: Yes or No.

Creation Motivation *

What motivated the creation of the database? Include statements in the paper motivating the creation of a new database.

Previous Databases

Comparison to Previous Databases

If the authors compare their database to any previous databases, list them here. Separate with semicolons (;) for ease of parsing. Use N/A if they do not compare to any previous databases.

How do authors describe this database's relation to previous databases? *

If the authors compared to any previous databases, paste the statements from the documentation into this column. Use N/A if they did not compare to any previous databases.

Limitations

Technical Limitations

Do the authors include limitations/challenges of the database itself?

Whether the authors included the limitations or challenges faced in creating the database itself (not related to models or validation). This relates to technical challenges of the database, such as gathering the data.

Options: Yes/No.

Link to the contract or terms of use *

If there is a link to a contract or terms of use, paste it here.

If the database has limitations/challenges, what are they? ✶

If the answer to the previous question was yes, how do the authors describe the limitations/challenges to the database?

If the answer to the previous question was no, write N/A.

Ethical Limitations

Do the authors explain or enact any explicit privacy considerations (of either collection or or use of the database)? □

Whether the authors included privacy considerations of the database explicitly. This can include privacy consideration in collection of the data or of use of the data. For example, whether the authors decided not to collect data for privacy or reasons, or limit certain uses of the data for privacy reasons.

For example:

- Any statements of privacy of the participants
- Any statements of privacy of annotators
- Any statements explicitly stating faces were blurred for privacy reasons

Options: Yes or No.

Do the authors explain or enact any explicit ethical considerations (of either collection or or use of the database)? □

Whether the authors have a specific explanation or section regarding ethical trade offs.

Options: Yes or No.

If there are ethical or privacy considerations, what are they? ✶

If the answer to the previous question was yes, how do the authors describe the ethical considerations of the database?

If the answer to the previous question was no, write N/A.

Usage Limitations

Did the authors describe the limitations of use? □

Whether the authors include some limitation to how the database can be used. This is more often on the website than in the research paper. This could include a license database users must fill out or a statement outlining the terms of service. Include both examples which require a license before accessing the database and which do not (e.g., there is a terms of use, but you can still download the database without explicitly agreeing to it).

Options: Yes, No, or N/A (when the database cannot be found).

If the authors describe limitations of use, what are they? ✶

If the answer to the previous question was yes, how do the authors describe the limitations of use? Copy the text from the website, paper, or license.

If the answer to the previous question was No or N/A, write N/A.

(If this is a separate form, enter the link in the "Documents Analyzed" column.)

Data Collection

Who Collected the Data

At a high level, do the authors give any information about where the data come from? □

Whether the authors state where the data for the database came from, whether it is from studio data collection, web scraping, etc.

Options: Yes or No.

Do the authors name who or what organization collected the data? □

Whether the authors describe who actually collected the data for the database. This could be the authors, some other organization, participants, or crowdworkers. Often, when it is the authors, it is implicit (e.g., “we collected…”).

Options: Yes or No.

Who collected the data thematic ✶

If the answer to the previous question was Yes, who collected the data?

If the answer to the previous question was No, write N/A.

Data Collection Type 1: Subset of Existing Database

Did the data come from a previous database or a subset of a previous database?

Whether the data for the database came from previous database(s). For example, for UMass FDDB, some images were taken from Faces in the Wild.

Options: Yes or No or N/A.

N/A should ONLY be used when the question “At a high level, do the authors give any information about where the data come from?” was answered No.

If the data came from another database, was it all or in part?

If the answer to the previous question was Yes, indicate whether the data that was taken from a previous database was All or In Part.

If the answer to the previous question was No or N/A, put N/A.

If the data came from another database, what database?

If the answer to the first question in this section was Yes, list the databases the previous data came from. Separate database names with semicolons (;) for ease of parsing.

If the answer to the first question in this section was No or N/A, put N/A.

If the data came from another database, did the authors state they got permission explicitly?

If the answer to the first question in this section was Yes, did the authors state anywhere that they got permission to use the data for a new database? This may be listed in the Acknowledgements section. Options: Yes or No.

If the answer to the first question in this section was No or N/A, put N/A.

If they got permission, what did they write?✶

If the answer to the previous question was Yes, paste what they wrote about getting permission.

If the answer to the previous question or the first question in this section was No or N/A, put N/A.

Data Collection Type 2: Studio

Was the data collected in a controlled environment, like a studio? □

Whether the data was collected in a studio environment with controlled variables like lighting, sound, etc.

Options: Yes or No or N/A.

N/A should ONLY be used when the question “At a high level, do the authors give any information about where the data come from?” was answered No.

Yes for synthetically created data.

If the data was collected in a controlled studio environment, did authors describe how they recruited subjects? □

Whether the authors describe how they recruited human subjects for studio data collection.

Options: Yes or No or N/A.

If the answer to the previous question was No, put N/A.

If the authors did not use human subjects, put N/A.

If the authors described recruitment, what did they do?✶

If the answer to the previous question was Yes, how did they describe recruiting human subjects?

If the answer to the previous question was No or N/A, put N/A.

Data Collection Type 3: Public Websites

Was the data collected from public websites or other public data? □

Whether the data for the database came from public websites, like Flickr.

Options: Yes or No or N/A.

N/A should ONLY be used when the question “At a high level, do the authors give any information about where the data come from?” was answered No.

If from public websites, did the authors discuss the original data license? □

Whether the authors discuss the original licenses of the data collected from public websites.

Options: Yes or No or N/A.

If the answer to the previous question was No or N/A, put N/A.

If they discussed the data license, how? ✶

If the answer to the previous question was Yes, how did they describe the previous license?

If the answer to the previous question was No or N/A, put N/A.

Data Collection Type 4: Real World Data Collection

Was the data collected by authors or participants in real world public settings? □

N/A should ONLY be used when the question “At a high level, do the authors give any information about where the data come from?” was answered No.

Human Subject Consent

Do the authors state whether those featured in the database can opt-out post collection? □

Whether the authors state in either the paper or other documents that human subjects can request to opt out of being featured in the database.

Options: Yes, No, or N/A.

N/A should only be used when there are only non-human subjects in the database.

Do the authors state whether consent was given by the subjects?

Whether the authors explicitly describe getting consent from human subjects.

Options: Yes, No, or N/A

Studio is not consent unless the word "consent" is explicitly used - “participant” and “participate” do not count towards consent.

N/A should only be used when there are only non-human subjects in the database.

Do the authors state whether informed consent was given?

Whether the authors sought *informed* consent from human subjects: the authors explained the database, the implications of being in the database, etc.

Options: Yes, No, or N/A

N/A should only be used when there are only non-human subjects in the database.

Do the authors state whether the subjects were compensated?

Whether the authors describe compensating human subjects directly for their data in some way. If the authors compensate only copyright holders who aren’t in the images, it does not count as compensating the subjects. This could include paying human subjects in cash, gift cards, credits, etc.

Options: Yes, No, or N/A

N/A should only be used when there are only non-human subjects in the database.

Do the authors state whether the study went through an IRB process (or international equivalent)?

Whether the authors describe the study going through an IRB or approval process before being conducted.

Options: Yes, No, or N/A

N/A should only be used when there are only non-human subjects in the database and no human annotators are involved.

Thematic Data Collection ✶

Paste any documentation about the data collection process, including any contextual information about the questions in this section.

Data Description

Does the database contain data of real humans? □

Whether the database contains images / videos / scans of real, existing human subjects.

Options: Yes or No

Does this contain images of synthetic humans? □

Whether the database contains images / videos / scans of synthetic (computer generated) human subjects.

Options: Yes or No

Are humans the main subject of the database? □

Whether humans are the main subjects of the database. If humans are in the database, but are not the main subject or purpose of the database, then the answer is No. For example, ImageNet contains human subjects, but its overall purpose is broad object recognition and thus it also contains many images of objects, animals, and scenes.

Options: Yes or No

Are real human faces visible and plausibly identifiable to the human eye? □

Options: Yes, No, or N/A

N/A should only be used when there are only non-human subjects in the database.

Demographics

Do the authors describe demographic categories of subjects? □

Whether the authors describe any demographic information about subjects. This may include gender, age, ethnicity, race, or a proxy for any of these.

Options: Yes, No, or N/A

N/A should only be used when there are only non-human subjects in the database.

If the data contains demographic information, what is it? ✶

If the authors include demographic information, how they describe it in the documentation. Paste quotes from the documents in this column.

Use N/A if the previous question was No or N/A.

Thematic Data Description ✶

Paste any documentation describing the data, including any contextual information about the questions in this section.

Annotations

Annotation Categories

Do the authors describe what kind of labels and/or ground truth are used in the paper? □

What kind of ground truth labels were defined in the paper regardless of whether they were annotated in the studio, by humans, or by machines). This might be facial expressions, subject ID, or metadata.

Options: Yes or No.

What types of labels and/or ground truth are attached to data instances? ✶

Describe the categories used in the database (regardless of whether they were annotated in the studio, by humans, or by machines). Separate by semicolons (;) for ease of parsing.

N/A if unknown or indeterminate.

Do data instances contain categorical labels? □

Whether the data in the database has categorical labels, such as types of objects, gender or race categories, etc.

Options: Yes or No.

If this data has categorical labels, is it necessary to download or load the data to understand the categorical schema/variables? □

Whether one has to download the database or load it as a package in a programming language to understand what kinds of categorical variables are used in the schema. For example, if the database contains labels for types of fish, does one have to download the database to know what kinds of fish?

Options: Yes, No, or N/A.

N/A should only be used when the answer to the previous question was No.

Where are the categories available? ✶

If the categories were described in the paper, where are they available? For example, are they available in the paper, some other paper, on the website, or if you have to download the database?

Categories Thematic ✶

Paste any descriptions of the categories from the documents into the column.

Post-Data Collection Annotations

Did the annotations come from a post-data collection annotation process?

Whether the data annotations came from a post-data collection process, rather than from data collection procedures. Annotations that came from data collection would include categories defined by studio data collection or by search queries. Annotations that come from post-data collection processes would only be assigned after data collection.

Options: Yes or No.

Yes = Human or Machine

No = Data collection as annotation

If there was a post-data collection annotation process, did the annotations come from Humans or Machines?

Who did the annotations for the annotation process. Machine annotation would be examples of automatically generated or derived labels.

If the process is semi-automated, make a determination based on the level of human involvement (e.g., just quality checking by authors = Machines vs. triangulation with crowd workers = Humans).

Options: Humans or Machines or Both or N/A

Both should be used when both Human and Machine annotation have been used.

N/A should only be used if the answer to the previous question was No.

Annotation Quality

Did the authors describe how they assessed annotation ground truth quality or disagreement?

Whether the authors describe assessing the quality of annotations or resolving disagreements.

Options: Yes, No, or N/A.

N/A if there are no human-generated annotations.

If the authors stated how they assessed annotation ground truth quality, what was the method? ✶

If the authors described assessing the quality of annotations or resolving disagreements, how did they do so? (e.g., interrater reliability)

N/A if the answer to the previous question was No or N/A.

Human Annotators

All of the following questions should be answered with Yes or No if the answer to “If there was a post-data collection annotation process, did the annotations come from Humans or Machines?” was Humans.

Did the authors describe who the annotators were? □

Whether the authors describe who the human annotators were for the database. For example, they may be Amazon Mechanical Turkers or students.

Options: Yes, No, or N/A

N/A should only be used if the answer to “If there was a post-data collection annotation process, did the annotations come from Humans or Machines?” was Machines or N/A.

Were the annotators the authors, students (who were not authors), third parties (like AMTs), or a mix? □

The kind of human annotator who did the annotations. Students indicates annotators who were not authors but were hired or recruited to do the annotation process separately. Third party annotators are any annotators who were neither authors or students, such as Amazon Mechanical Turkers, survey respondents, or FACS coders.

Options: Authors, Students, Third Party, or N/A

If the answer to the previous question was No or N/A, write N/A.

If the authors describe the annotators, who are they? ✶

If the answer to the previous question was Yes, write who the humans annotators were (e.g., Amazon Mechanical Turkers, college undergraduates, etc.).

Write N/A if the answer to the previous question was No or N/A.

Did the authors describe where annotators were recruited from? □

Whether the authors describe where they recruited the human annotators from. For example, from Amazon Mechanical Turk or from social media.

Options: Yes, No, or N/A

N/A should only be used if:

- the answer to “If there was a post-data collection annotation process, did the annotations come from Humans or Machines?” was Machines or N/A.
- the authors were the annotators.

If the authors describe where annotators were recruited from, where? ✶

If the answer to the previous question was Yes, write where the humans annotators were recruited from.

Write N/A if the answer to the previous question was No or N/A.

Did the authors describe how annotators were recruited (e.g., recruitment materials)? □

Whether the authors discuss the types of recruitment materials used to recruit annotators (e.g., surveys, flyers, etc.).

Options: Yes, No, or N/A

N/A should only be used if:

- the answer to “If there was a post-data collection annotation process, did the annotations come from Humans or Machines?” was Machines or N/A.
- the authors were the annotators.

If the authors describe the recruitment process, how were annotators recruited? ✶

If the answer to the previous question was Yes, write where the humans annotators were recruited from.

Write N/A if the answer to the previous question was No or N/A.

Do the authors describe if annotators were trained? □

Whether the authors state that the annotators were trained for the annotation task.

Options: Yes, No, or N/A

N/A should only be used if the answer to “If there was a post-data collection annotation process, did the annotations come from Humans or Machines?” was Machines or N/A.

If the authors state annotators were trained, how were they trained? ✶

If the authors state that the annotators were trained, paste any documentation on how they were trained in this column.

Write N/A if the answer to the previous question was No or N/A.

Did the authors state if the annotators were compensated? □

Whether the authors state that the annotators were compensated for their annotation labor (e.g., cash, gift cards, course credits, etc.).

Options: Yes, No, or N/A

N/A should only be used if:

- the answer to “If there was a post-data collection annotation process, did the annotations come from Humans or Machines?” was Machines or N/A.
- the authors were the annotators.

Did the authors include any demographic information on annotators? □

Whether the authors include any demographic information about the human annotators. This may include gender, age, ethnicity, race, or a proxy for any of these.

Options: Yes, No, or N/A

N/A should only be used if the answer to “If there was a post-data collection annotation process, did the annotations come from Humans or Machines?” was Machines or N/A.

If there is demographic info, what is it? ✶

If the authors include demographic information, how they describe it in the documentation. Paste quotes from the documents in this column.

Use N/A if the previous question was No or N/A.

Did the authors state whether annotators were given or the authors developed either annotation procedures, a codebook, or instructions for annotation? □

Whether the annotators were given instructions or a codebook for annotating. If the annotators were the authors, whether they developed a codebook.

Options: Yes, No, or N/A

N/A should only be used if the answer to “If there was a post-data collection annotation process, did the annotations come from Humans or Machines?” was Machines or N/A.

If the annotators were given or developed a codebook, is that codebook publicly available? □

Whether the authors published their annotation codebook or instructions.

Use N/A if the previous question was No or N/A.

Thematic information on annotations ✶

Paste any documentation describing the human annotation process, including any contextual information about the questions in this section.

Values

Values: Values & Value-Laden Terms ✶

How the database authors valued specific aspects of the database. Which words are imbued with value, and are those words described or do they need to be inferred by the readers (e.g., “challenging”)? Keep a log of these statements in the spreadsheet.

- Examples
    - ImageNet, Paper2:
        - "accuracy" & "state-of-the-art" : This paper describes the creation of this benchmark dataset and the advances in object recognition that have been possible as a result. We discuss the challenges of collecting large-scale ground truth annotation, highlight key breakthroughs in categorical object recognition, provide a detailed analysis of the current state of the field of large-scale image classification and object detection, and compare the

state-of-the-art computer vision accuracy with human accuracy." & ILSVRC makes extensive use of Amazon Mechanical Turk to obtain accurate annotations" & "To collect a highly accurate dataset, we rely on humans to verify each candidate image collected in the previous step for a given synset. "

- Non-examples
    - H3D: An attitude or posture of the body, or of a part of the body, esp. one deliberately assumed, or in which a figure is placed for effect, or for artistic purposes. This definition captures the two aspects of a pose: A configuration of body parts such as head, torso, arms and legs arranged in 3D space. The resulting appearance, a 2D image created for a viewer, or a camera.

Values: Are the values in the paper explicitly defined or assumed? □

Do the authors define what they mean when using value-laden terms (e.g., quality, clean, challenging) or assume the reader understands the meaning?

Options: Explicit, Assumed, or Mixed

Documentation

Did they name the database in the original publication? □

Whether the authors gave the database a name.

Options: Yes or No

If the authors name the database, what did they name it? ✶

The name the authors gave the database.

Use N/A if the previous question was No.

Did they give the data a DOI? □

Whether the authors gave the data a stable DOI.

Options: Yes or No

Website Documentation

Website cited in paper ✶

If there was a website in the original paper, put the links here.

Use N/A if there is no website in the paper.

Website currently in use (if different from website cited in paper) ✶

If there is a new website that is being used to host the database, paste the links here.

If the website currently in use is the same, use N/A.

Also use N/A if there is no current website.

Is there a website cited in the paper for the database? □

Whether the authors provide a link to where the database is hosted in the original paper.

Options: Yes or No

Is the website still available? □

Whether the website in the original paper is still available.

Options: Yes, No, or N/A

N/A when the answer to the previous question was No.

Is there a website for the database at all? □

Whether a website hosting the database can be found at all.

Options: Yes or No

Checked website availability (Date) □

Date the researcher checked the website availability.

Did they host the database on a personal website/account and/or lab website? □

Whether the database was hosted on a personal or institutional website with no stable DOI.

Options: Yes, No, or N/A

N/A if no website is available.

Did they host the database on a data repository (e.g., Zenodo, Dataverse)? □

Whether the database is hosted in a data repository with a stable DOI.

Options: Yes, No, or N/A

N/A if the database is not available.

Availability

Is the database still available for download? □

Whether the database is still available for download anywhere.

Options: Yes, No, or N/A

N/A if the database requires registration/approval or package installation.

Checked database availability (Date) □

Date the researcher checked the database availability.

Database Availability Thematic ✶

Any additional information about the database availability (e.g., link in original paper is dead, only partially available, etc.).

Paragraph Proportions

Where is the database most documented? ✶

Whether the database is most documented in the paper, on the website, or elsewhere.

Where in the paper is the database documented? ✶

Which sections of the paper contain the most content documenting the database.

Paragraphs Dedicated to Database Documentation □

Number of paragraphs that give any contextual information about the database itself. This may be only one sentence in a paragraph (including funding in the acknowledgments), or an entire paragraph.

Substance of paragraphs to be included

- Prior work (comparison to prior datasets)
- Data collection, description, annotation
- How to use the dataset

Total Paragraphs □

Number of paragraphs in the entire paper.

Methodology for counting paragraphs

- Lumps paragraphs with two lines or less into the next paragraph
    - If there is a hanging word or two on a two line paragraph, count it as two lines.
- Lump equations within the paragraph they are discussed in
- For bulleted lists:
    - If the bullet points are one line each, count the entire list as one paragraph
    - If each bullet point is more than three lines, count the bullet points as separate paragraphs
- Do not count figures or figure captions, even if they are more than two lines
- Do count abstracts and acknowledgements.
- Count appendix only if more than two lines of text are dedicated.

Proportion of Paper Dedicated to Database □

The proportion of content in the paper dedicated to the database. Paragraph count divided by paragraph total.

Time Spent (in Minutes) □

The amount of time it took to code the information in the spreadsheet. When updating information on a database, remember to also record the amount of time spent and sum all time. Round down to the minute for 29 seconds and under; round up to the minute for 30 seconds and over.