Published in Proceedings of SUMAC 2020, online (originally should have taken place in Seattle, USA), 12 October 2020, which should be cited to refer to this work.
DOI: 10.1145/3423323.3423413

Hybrid Human-Machine Classification System for Cultural Heritage Data

Shaban Shabani[1,2]
shaban.shabani@unibas.ch

Maria Sokhn[2]
maria.sokhn@hes-so.ch

Heiko Schuldt[1]
heiko.schuldt@unibas.ch

[1] Mathematics & Computer Science
University of Basel
Basel, Switzerland

[2] University of Applied Sciences
Western Switzerland (HES-SO)
Neuchâtel, Switzerland

ABSTRACT

The advancement of digital technologies has helped cultural heritage organizations to digitize their data collections and improve the accessibility via online platforms. These platforms have enabled citizens to contribute to the process of digital preservation of cultural heritage by sharing documents and their knowledge. However, many historical datasets have problems due to incomplete metadata. To solve this issue, cultural heritage organizations heavily depend on domain experts. In this paper, we address the issue of completing the metadata of historical digital collections. For this, we introduce a new hybrid human-machine model. This model jointly integrates predictions of a deep multi-input model and inferred labels from multiple crowd judgements. The multi-input model uses visual features extracted from the images and textual features from the metadata, complemented with Wikipedia classes of concepts extracted in the text. On the crowd answer aggregation, our method considers the workers' reliability scores. This score is based on the performance of workers' task history and their performance in our task. We have applied our hybrid approach to a culture heritage platform and the evaluations show that it outperforms both deep learning and crowdsourcing when applied individually.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; • **Information systems** → **Crowdsourcing**.

KEYWORDS

Cultural heritage; deep learning; crowdsourcing; hybrid human-machine information systems

ACM Reference Format:
Shaban Shabani[1,2], Maria Sokhn[2], and Heiko Schuldt[1]. 2020. Hybrid Human-Machine Classification System for Cultural Heritage Data. In *2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC'20), October 12, 2020, Seattle, WA, USA*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3423323.3423413

SUMAC'20, October 12, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8155-0/20/10…$15.00
https://doi.org/10.1145/3423323.3423413

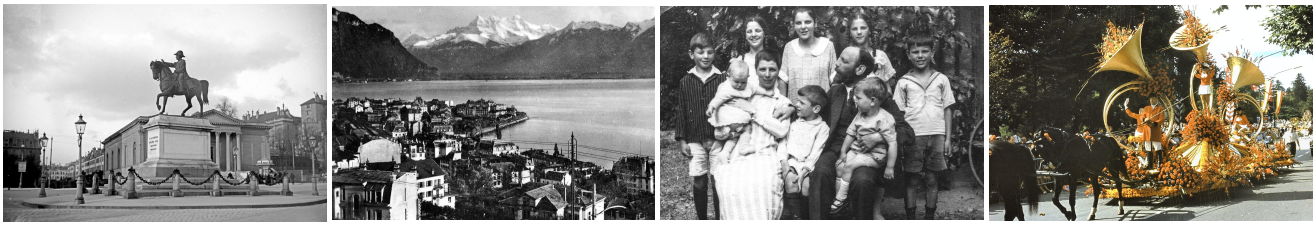
1 INTRODUCTION

Citizens and cultural heritage institutions have crucial roles to play for transferring historical information between generations and civilizations. Many citizens own valuable data, such as photo albums, or audio and video archives, that can be of great public interest. Considering the rapid advances in mobile internet technology, many initiatives motivate digitizing and sharing collections via online platforms. However, platforms providing reliable methods for the documentation and management of cultural and historical data only partly solve the problem since digitized content is unusable without proper metadata. Therefore, these digital management platforms produce metadata that are of utmost importance. They cover the spatio-temporal properties of the data (date and location), the descriptive aspects of items (title, description, tags, and category), and their provenance properties (creator, owner, and license). However, obtaining complete metadata for any item poses several challenges.

Numerous historical datasets are poorly annotated, for instance, as images may be missing tags and are not categorized. This is mainly due to annotators' mistakes or lack of accurate information. Incomplete cultural data lose their relevance in search engines and consequently their historical value when they cannot be found. Classifying and annotating historical data are tedious tasks, and cultural heritage institutions usually ask professionals to provide high-quality annotations. However, due to large data collections, doing in-house annotations with different specialists who have different levels of expertise in various domains is difficult and does not scale.

Considering the constraints mentioned above, many cultural heritage institutions such as Galleries, Libraries, Archives and Museums (GLAMs) are more and more exploring the potential of crowdsourcing [24]. These institutions make use of the knowledge and capacity of the crowd [21], by opening their collections and by inviting online users to contribute by annotating data. For instance, the Australian Newspaper initiative from the National Library of Australia [15] opened their inaccurately digitized newspapers and invited online volunteers to correct the wrongly optical character recognized (OCR) text. Another project by Steve.Museum [35] invited online contributors to tag works of art from the museum, and tags were later compared to the museum documentation. The results showed that big proportions of tags represented terms not found in museum records. These initiatives open up collections to the online crowd to enrich data collections.

Nevertheless, the quality of crowdsourced data remains a challenge to be addressed. Compared to domain experts, who follow strict guidelines when annotating the data, crowd users are not



(a) place, object (b) place (c) person(s) (d) event, tradition

Figure 1: Sample images from the dataset with the corresponding annotated categories. Figures (a) and (d) have two categories whereas Figures (b) and (c) have single categories.

trained and data quality is not guaranteed. Depending on the sensitivity of the data, crowdsourced annotations need an additional assessment. In some scenarios, automatic quality mechanisms such as qualification tests and aggregation mechanisms [1] would be enough to maintain a high quality of data. In other cases, it is desirable to evaluate the manual assessment of crowdsourced data by domain experts if quality criteria suffer. However, quality control affects the latency and cost of the annotation process.

With the rapid growth of online data, having effective and efficient data processing tools is essential. Machine learning based systems have been used and comprehensively applied to solve various labeling tasks, such as image classification [11], object detection [13], and sentiment annotation [37]. Recent advances in machine learning and deep learning have increased the interests of developing tools to annotate and classify cultural heritage data as well [2, 17]. However, automated tools fall short of performing as accurately as humans could. To tackle the three-dimensional problem of accuracy, latency, and cost, hybrid human-machine systems [8] have been proposed. Hybrid approaches are highly promising since they leverage the scalability of machines over a large amount of data and the quality of human intelligence. Such systems are meant to combine the efficiency of computer algorithms with the wisdom of crowds [32].

In this paper, we propose and evaluate a hybrid human-machine multi-input approach for historical data classification in the context of GLAMs. The classification task focuses on categorizing images with historical content. Following the guidelines [4] of categorizing cultural heritage items, the goal of this categorization task is to enrich the metadata of the dataset by adding a new attribute, "cultural_interest". This is done by assigning each image to one or more of the following five categories: *place*, *object*, *person*, *event*, or *tradition* [23].

For this multi-label classification task, we initially consider a hybrid multi-input transfer learning approach to automatically classify the images. This approach combines a deep learning model pre-trained on image visual features with a model that explores text features from the text available in the metadata. The results show that the hybrid model performs better than the individual models. Nevertheless, we consider these results imperfect. Therefore, for the same multi-label classification task, we ask online crowd workers from a crowdsourcing platform to also categorize the images. Finally, we implement and evaluate the hybrid human-machine approach as an advantageous solution. This approach provides higher accuracy at an acceptable cost and latency. It combines the effectiveness of deep learning algorithms with the wisdom of crowds, through the application of: i.) high-confidence switching, i.e., whenever machine learning algorithms fail to categorize the images with high accuracy, human feedback is used in the annotation process; ii.) an aggregation model where the judgments of crowd users and machine-based model are aggregated with the aim of increasing higher quality of annotations.

Our contributions can be summarized as follows:

- We implement and evaluate a deep learning multi-input approach by combining pre-trained models that leverage the image visual features that explores the text features available in the metadata.
- We design a crowdsourcing task that leverages the human cognitive skills of online crowd workers to categorize historical images. We implement a customized aggregation model that considers the crowd users' profiles for the answer selection.
- We implement and evaluate two hybrid human-machine approaches for image classification. The hybrid approaches perform better than automatic classification and crowdsourcing schemes.

The remainder of the paper is structured as follows: Section 2 introduces the data and describes the individual deep learning and crowdsourcing methods, as well as our hybrid human-machine system for cultural heritage data categorization. Section 3 details the evaluation and results of our methods. Section 4 presents related work and Section 5 concludes.

2 DATA AND METHODS

In this section, initially we present the dataset. After that, we describe the overall architecture and the tools and methods used to design and implement the human-machine image classification approach.

2.1 Dataset

The dataset was taken from NotreHistoire[1], which is a participatory platform for sharing and valorizing the history and cultural heritage of different regions of Switzerland. This platform invites volunteers to publish and share their own digitized archives along with institutions (e.g., galleries, libraries, archives, museums, radio

[1] https://www.notrehistoire.ch

Figure 2: Web annotation tool

and television broadcasters, etc.). The platform's database mainly contains images, as well as videos, audios, and documents. The data selection process included further inspection of licenses of shared data, filtering only images that had a "by-nc-nd" Creative Commons license[2]. This is important in order to reproduce the experiments and validate the results we obtained. Finally, the dataset contains 5,015 images and metadata information about the images such as title, description, location, year, tags, and author information. Figure 1 illustrates some example images from the dataset.

2.2 Image annotation

The aim of this work is to enrich the metadata of the dataset with a new attribute *cultural_interest* by assigning each image one or more of the following categories:

(1) *place* – if the image is showing a place/location (e.g., landscape, mountain, city view)
(2) *person* – if the main theme of the image is a person or group of people and people are clearly identified in the image (e.g., portrait of people)
(3) *event* – if the image depicts an organized event (e.g., carnival, festival)
(4) *object* – if the image shows an object (e.g., sculpture, painting, specific vehicle or building)
(5) *tradition* – if the image shows people with specific clothes in events or performing particular activities

Since the dataset was not labelled, ground-truth data was needed in order to evaluate the performance of automatic classification, crowdsourcing, and hybrid human-machine approaches. For this reason, we organized an annotation task by inviting 10 participants from the region. To participants were presented the idea behind the project and the objectives. They were shown detailed instructions with comprehensive examples on how to annotate the image. In two rounds, the 10 trained annotators used a developed web annotation interface tool to annotate the images (Figure 2). In the first round, they used a pooling mechanism where each annotator was assigned an image that had not been annotated. In the second round, annotators were assigned images that had already been annotated by other annotators from the first round. Inter-rater agreement was $\kappa = 0.55$, and considering that it is a multi-label task, this number can be considered substantial [19]. An additional, third round was required to resolve annotators' disagreements. Table 1 represents the number of images that contain each of the five categories. Clearly, the most frequent category in the images is *place* which appears in 63% of the images, followed by the categories *object* and *person* with 35% and 22%. The categories *event* and *tradition* are least represented categories in the dataset with 10% and 5%, respectively. Table 2 shows the number of categories per image. A single category appears in 3,382 or 67% of images, 1,502 images or about 30% have two categories, 129 images or about 3% contain three categories, only 2 images result to have four categories, and there was no image that had all five categories.

[2]https://creativecommons.org/licenses/by-nc-nd/2.0/

Table 1: Number of images that contain each category

Category	Number of images	Percentage
place	3,197	63%
object	1,742	35%
person	1,098	22%
event	515	10%
tradition	229	5%

2.3 Automatic classification via transfer learning

Thanks to the high numbers of images available online, image classification and object detection are some of the areas where deep learning has shown promising results. A deep neural network trained on enough large dataset can classify images with high accuracy. For instance, the ImageNet project [9] has a very large database of 14 million hand-annotated images that contains more than 20,000 classes. However, the process of collecting and annotating a dataset is expensive and time-consuming. Moreover, developing a new model from the ground up every time on small training data does not provide high accuracy. Therefore, pre-trained models and transfer learning [27] techniques reduce the effort needed to collect massive amounts of training data.

Pre-trained models are trained in a context of large and general classification tasks. Therefore, they can be used to address a more specific task by extracting and transferring meaningful features that were previously learned. Some of the popular image pre-trained models are VGG19[33], MobileNet[16], and ResNe[34]. We use the MobileNet deep neural network and implement transfer learning with a fine-tuning method. The customized implementation has a logistic regression final layer with sigmoid activation function and uses the binary cross entropy loss [26]. The last predicting layer

Table 2: Distribution of categories over the image dataset

#categories	one	two	three	four	five
#images	3,382	1,502	129	2	0
#percentage	67.44%	29.95%	2.57%	0.04%	0%

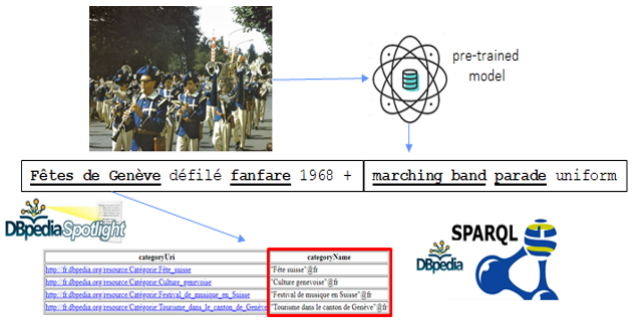


Figure 3: Extracting concepts and DBpedia categories

of the pre-trained model is removed and replaced with the custom predicting layer that contains the five categories: person, object, place, event, and tradition. The multi-label classification model for each of the images assigns an image to one or more classes.

In parallel to building a model that relies on visual features, we investigate the importance of text information available in the existing metadata attributes. Our method uses an image labeling model to get image tags, entity extraction to find concepts, and DBpedia knowledge base to query concept categories. Our assumption is that the DBpedia categories can provide more specific information that connect directly with the image category we aim to predict. This is better illustrated in Figure 3. The example image is taken during the *Geneva Festival* and the title of the image is *Fetes de Geneve defile fanfare 1968*. The pre-trained vgg model additionally provides the tags "marching band", "parade uniform" in English. The entity extraction tool recognizes the concepts "Fetes de Geneve", "fanfare", "marching band" and "parade". Exploring the DBpedia category for the "Fetes de Geneve" concept gives us the "Fete Suisse" which is another event. On the other hand, the categories of "marching band" and "parade" give more context that this event is a tradition as well.

2.4 Crowdsourcing approach

Collecting annotated data is an expensive and time-consuming process. Crowdsourcing has been widely used as alternative service to replace experts with specific domain knowledge for labelling. It efficiently reduces the costs and latency by making use of the collective intelligence of thousands of available crowd users on the Internet. There are many popular non-paid crowdsourcing projects in citizen science [14] such as Wikipedia, GalaxyZoo [5], and Recaptcha [38]. In parallel, there are several popular commercial crowdsourcing online platforms such as Amazon Mechanical Turk (MTurk)[3], FigureEight[4], MicroWorkers[5], etc. These platforms enable the exchange of HITs (Human Intelligence Tasks) between requesters who need tasks to be completed, and workers who are available and willing to complete a task, and who get a financial reward for that work.

For categorizing the images of the dataset, we used the MicroWorkers crowdsourcing platform. Each image was used to generate a HIT, asking online crowd participants to categorize the image into one or more of the 5 classes. A HIT contained the URL of the image. Additionally, we added the title, location and description of the image that could provide some helpful context. Crowd workers were instructed to analyze the image and provide the most suitable classes for that image. The five classification categories were place, object, person, event, or tradition. As task design techniques [12] are important factors that increase the quality of crowdsourced data, guidelines with tips and examples were therefore part of the instructions. Furthermore, to ensure quality control, the golden question [22] technique was applied. We used a set of qualification tasks to allow only qualified users who pass the test with an accuracy of at least 60% to keep on labelling the images. Additionally, reputation mechanisms [6] were enabled, opening the annotation job only to the *best microworker group* that has the workers with the highest reputation on the platform. Asking multiple crowd workers to perform the same task is used usually to increase the quality of the data by aggregating the answers. Therefore, for each image we asked three crowd users to provide the answer. Depending on the task, sometime even simple Majority Voting (MV) aggregation algorithm increases the data quality.

Considering that in our case we have a multi-label task, having multiple judgments can lead to higher disagreement between annotators, yielding to low-quality answers. Therefore, we applied three truth inference algorithms to infer the correct answer from the workers' answers. We decompose the worker answers into binary form. For instance, if a worker answer is *place* and *object*, from the set of classes ["place"", "object", "person", "event", "tradition"], the answer is encoded into a binary vector $[1, 1, 0, 0, 0]$. First, we consider MV algorithm which simply select as final the answer given by the majority of workers. Next, we evaluate the Dawid and Skene model [7] which is based on the Expectation-Maximization (EM) principle to model the worker's reliability with a confusion matrix for the answer aggregation. Last, we apply a truth inference algorithm that considers prior information provided by the crowdsourcing platform about worker's profile. We derive a reputation of a worker based on his previous finished tasks (number of accepted and rejected tasks, money, and badges earned) and integrate that score in a weighted aggregation method to infer the true answer.

2.5 Hybrid human-machine annotation

While crowdsourcing reduces the cost and latency per annotation compared to domain experts, hybrid human-machine information systems aim at reducing the overall annotation costs by selecting only the most important instances for being annotated by humans.

The *high confidence switching* as a hybrid method only selects instances for which the machine learning model is uncertain. Current machine learning models also provide a confidence estimate [30] on how accurate their answer is. Therefore, predictions with low confidence values are considered further to be solved by crowd workers. This method is helpful in scenarios when the availability of human annotators is limited and latency is critical as it tries to boost the accuracy of automated algorithms by minimizing the human input.

The *human-machine aggregation* is another hybrid method where the predictions of the machine learning model and the inferred crowd answers are jointly combined to resolve the final output.

[3]https://www.mturk.com/
[4]https://figure-eight.com/
[5]https://microworkers.com/

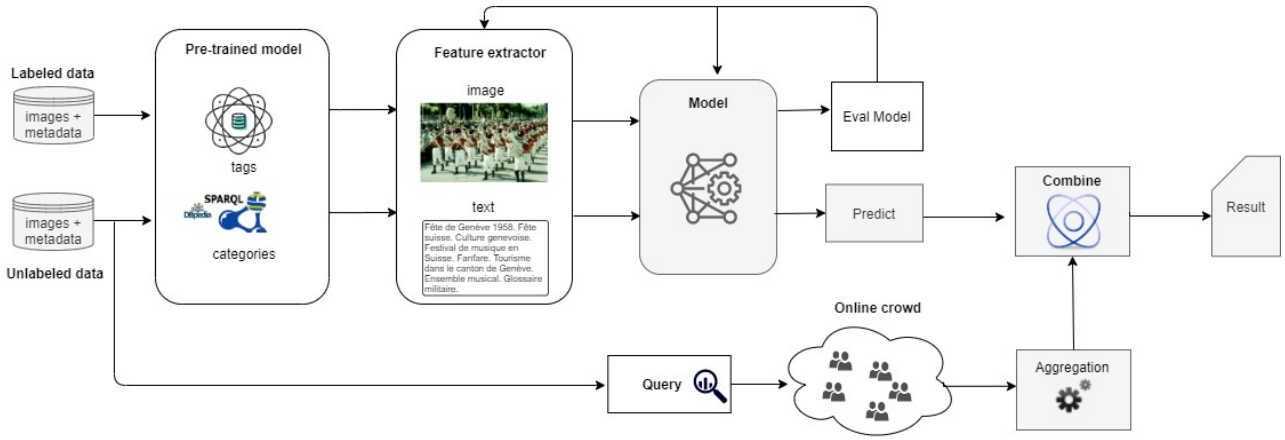


Figure 4: Overall architecture of the proposed hybrid human-machine approach

A multi-label classification task with multiple human annotators in some settings is prone to higher disagreement and a consensus on the output is not reached. On the other hand, for specific cases, machine learning models are not able to depict the context as good as humans. This method aims at combining weak responses to eventually increase the quality of results. The assumption is that the fusion in the aggregation will cancel eventual individual weaknesses. The *human-machine aggregation* method is suitable especially for scenarios when latency and cost of classification is not an issue but accuracy is essential. We experience this situation with the NotreHistoire platform. This project is running for several years and it has thousands of registered users and hundreds of active members. The members volunteer in sharing new historical content and they actively participate in the data curation process. We consider a weighted aggregation, where the estimates of the deep learning and the crowd aggregation models are multiplied with different scores. Their weights are derived based on their individual classification accuracies on the validation set. The sum of the weighted estimates results as the joint predicted output. The individual class estimates higher than a threshold are taken as predictions. Figure 4 illustrates the full pipeline of the human-machine approach for image categorization.

3 EVALUATION

In this section, we outline the evaluation of the proposed methods for our multi-label image categorization problem on the NotreHistoire dataset. In traditional binary and multi-class classification problems, commonly used evaluation metrics are precision, recall, and the F1 score. But in multi-label classification tasks, there are additional evaluation metrics such as exact match accuracy (subset accuracy) and Hamming-loss. Exact match is a strict metric measuring the percentage of the samples that have all their labels classified correctly, whereas the Hamming-loss measures only the fraction of wrong labels to the total number of labels, thus penalizing the individual labels. We use these two metrics to evaluate the performance of the automatic classification based on visual and textual features, the crowdsourcing approach, and the hybrid human-machine method.

3.1 Automatic classification

The starting point for the evaluation is to split the modeling data into training, validation, and testing sets. We decided to allocate 60% of the data for training, 20% for validating the models, and 20% for the test set. The original dataset classes are strings that are easy to understand by humans. However, to build and train a neural network model on a multi-label scenario, binary labels are generated from multi-hot encoding. Since the image dataset has images with metadata attached to them, we considered the available text from the metadata attributes: "title", "description", and "tags" as input text to the model. Additionally, from the concatenated text, we extracted labels with a pre-trained VGG16 model [33]; extracted the DBpedia concepts with dbpedia-spotlight [20]; and the DBpedia categories of each extracted concept were retrieved with the DBpedia SPARQL endpoint. Finally, each image metadata has the title, description, user provided tags (if available), automatic extracted labels, DBpedia concepts, and these concepts' categories.

Initially, we evaluate the accuracy of the machine learning models by considering only textual features. Feature extraction is run on the final combined text by using the term frequency, inverse document frequency (tf-idf) with the following parameters: $min\_df = 3$, $max\_features = 3000$, $stop\_words$ = English + French, $use\_idf = 1$, and $analyzer$ = word. After that, we selected three different machine learning classification models: i.) Logistic Regression (Log), ii.) Random Forest (RF), and iii.) Support Vector Machines (SVM). Table 3 presents the classification results based only on textual features. While accuracies of the three models are similar, the SVM model achieved the highest accuracy of 58% on the testing set.

Table 3: Accuracy of machine learning models according to textual features

Classifier	Accuracy	Hamming-loss
Logistic Regression	56%	13%
Random Forest	57%	12%
Support Vector Machines	58%	12%

Table 4: Accuracy of models with different combination of text data

Combination	Logistic	Random Forest	SVM
TD	49%	49%	51%
TDV	51%	52%	54%
TDVE	55%	56%	57%
TDVEC	**56%**	**57%**	**58%**

To analyze the importance of the generated features, we run the evaluation of the models separately based on different settings. We consider the following combinations of text available on the metadata and the additional generated data:

(i) title + description *(TD)*
(ii) title + description + vgg labels *(TDV)*
(iii) title + description + vgg labels + dbpedia-spotlight entities *(TDVE)*
(iv) title + description + vgg labels + dbpedia-spotlight entities + dbpedia categories *(TDVEC)*

Table 4 shows the accuracy of each model when using features extracted on the combinations mentioned above. Considering only the originally provided text on the metadata of the images (TD), SVM achieves an accuracy of 51% and a Hamming-loss of 12%, whereas the RF and Log models achieve an accuracy of 49% and Hamming-loss of 12%. Our initial assumptions that adding more information will boost the accuracy were valid, and this can be observed from the results shown in Table 4. All three models perform better when more text information is added to the input. The text provided by the TDVEC combination reached the highest accuracy of 58%.

Our next step was to evaluate the performance of a deep learning multi-input model that combines text and image features. Considering that our dataset consists of 5,015 samples, the size of this collection would not be large enough to build and train a deep learning model from scratch. Therefore, we apply transfer learning techniques by using the MobileNet [16] and GloVe [29] pre-trained models. MobileNet model was set to use weights from ImageNet which is trained on a large image collection and with a more general classification task. We configured it with a depth multiplier of 1.0 and an input size of 224×224. A new classifier with our custom dataset labels on top of it was added. Accordingly, the dataset images were resized to adapt to the input expected by the MobileNet model. In the text input, we added an embedding layer with loaded weights from the GloVe pre-trained word embeddings.

The new custom classification head was trained with images of our dataset (training set), so that the model addresses the multi-label classification task. A learning rate of 1e-5 on the training process was used, and the performance on the validation set was measured on 30 epochs. Figure 5 outlines the training and validation accuracy score of the multi-input deep learning approach. After 30 epochs, our model achieved an accuracy of 63% and Hamming-loss of 10% on the validation set. It is important to emphasize the effect that the transfer learning method has on the model's accuracy. During the first 10 epochs of training and the validation process, we set the layers of the MobileNet pre-trained model as non-trainable (frozen). After that, the last 100 layers (of the total 155 layers) were "unfrozen"

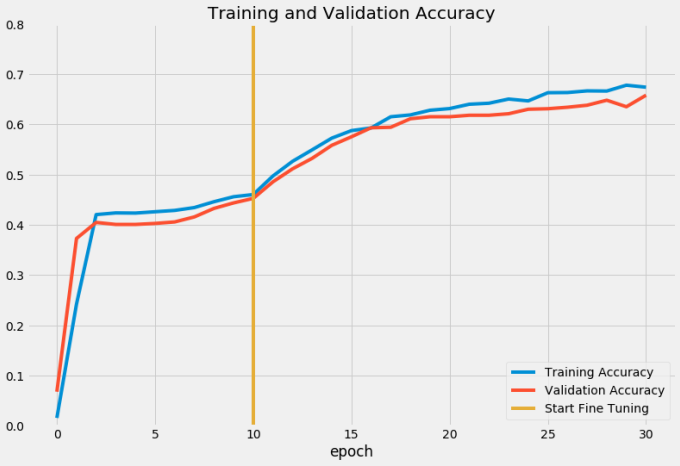


Figure 5: Training and validation accuracy on multi-input text and image model

and we retrained the model for another 20 epochs. We can observe that the accuracy increases (after the yellow vertical line) as an effect of transfer learning. Finally, we evaluated the model with data from the testing set and the accuracy achieved was 62% with a 10% Hamming-loss.

3.2 Crowdsourcing results

Considering that this is a multi-label task, we assume that lower represented classes such as "event" and "tradition"" are more challenging for a model to predict. In contrast to this, humans have the potential to perform better, especially when identifying if an image is from the context "tradition" or "event". As a result, crowdsourcing has been considered an alternative solution to this image categorization problem. For the same categorization task, we thus make use of the MicroWorkers crowdsourcing platform to generate HITs, asking online crowd workers to categorize each image from the dataset. The entire set of images was used to generate 5,015 HITs, one HIT per image. The original metadata of the image collections are in French and automatic classification models use this format. However, for the online crowdsourcing task, we translated the text automatically to English to expose the task to the largest crowd worker groups of the platform who speak English. Task design techniques [12] are essential factors that increase the quality of crowdsourced data. Therefore, elements such as instructions, rules, tips, and examples have been thoroughly considered in our project to guide and help the online workers to solve the tasks. Furthermore, a reputation-based method was applied, by opening the job only to "best annotators" group of the platform. For each image, we asked three workers to provide the categories. The annotation by online crowd users took 8 hours effective time to complete with 542 participants from 78 countries on average completing 31 tasks. Since our goal is to compare the accuracy of crowdsourcing with automated approach, we use the annotations from the validation set and testing set to compare the three different aggregation methods described in Section 2.4. Our assumption that in a multi-label task redundancy can lead to higher disagreement were confirmed. If we

Table 5: Results from crowdsourced data aggregation methods, deep learning, and hybrid human-machine method

Method	Accuracy	Hamming-Loss
#1 worker	49%	14%
#2 workers	47%	14%
#3 workers	41%	16%
Majority Voting	56%	12%
Dawid-Skene	58%	12%
Worker-Profile	**58%**	**11%**
Deep Learning	**62%**	**11%**
Hybrid human-machine	**65%**	**10%**

simply aggregate the answers without any quality control mechanism, adding the responses from the second and the third annotator reduce the accuracy from 49% to 41%. Data quality is a major issue in crowdsourcing, therefore we evaluate the MV algorithm, the Dawik-Skene model, and our worker profile model. Table 5 details the results obtained from the crowdsourcing experiment. The MV algorithm decomposes the task in binary output where for each of the five classes (place, object, person, event, tradition) the majority vote is taken as a final answer. The majority voting achieves an accuracy of 56% and hamming-loss of 12%, the Dawid-Skene and worker-profile model achieve slightly higher accuracy of 58% where the later one has lower hamming-loos of 11%.

3.3 Hybrid human-machine image categorization

So far, we have observed that text information from the image metadata, together with additional semantic information extracted from Wikipedia, can improve the classification accuracy of machine learning models. On the other hand, due to the multi-label task, disagreements between annotators resulted in lower accuracy of crowdsourced data. However, we expect that joining the outputs of the two approaches will improve the overall accuracy, complementing each other's strengths.

To evaluate the hybrid human-machine aggregation method described in Section 2.5, we use a weighted sum of the class estimates of deep learning model and the inferred classes from the answer aggregation of crowd workers. Since the automatic approach provided better results in general, we expect that weighting higher its output compared to the crowd answer will perform better. Therefore, the validation set was used to test different weight scores for the deep learning and crowd outputs. We found that the weights of 0.7 for deep learning and 0.3 for crowd outputs achieved the highest strict accuracy of 65% and hamming-loss of 10%. Incorporating the human judgments in the final output showed to improve the accuracy by 3%. Since the image categorization is a multi-label task with five classes, this can be considered as an improvement.

3.4 Discussion

Automatically classifying the images by methods that used visual features and textual features achieved an accuracy of 58% and 62%, respectively. The reason why the multi-input deep learning model performs better is that the information about the extracted concepts and their categories gathered from DBpedia extend the context of the image item. This extension is especially helpful for the classes "event" and "tradition" since it is difficult to extract features that can distinguish between these two classes for an image. Although the overall improvement of 4% in accuracy and 1% for the hamming-loss is not high, it is nevertheless higher for the two less represented classes of the dataset ("event" and "tradition"). Such an example is illustrated in Figure 3.

On the other hand, we experienced that the highest accuracy achieved by aggregating crowd answers was 58%, which is lower than the accuracy of the deep learning model – however with the hamming-loss being the same. One reason for the lower accuracy is that for specific images there is a disagreement between annotators, especially when judging whether the image class is "place" and/or "object". In the answer aggregation, we note that majority voting does not require prior data, whereas the Dawid-Skene model used the training set to estimate the workers' reliability. In the worker-profile method, we did not use the training set, however, we relied on profile information provided by the MicroWorkers platform.

4 RELATED WORK

Recent works focus more on automatic classification of historical data with deep learning techniques. These works cover image classification of different type of artworks such as paintings, statues, archeological artifacts, and architectural object designs. Llamas et al. [17] focus on classification with deep learning of images of architectural styles relying on visual features. Belhi et al. [2] apply a multimodal deep learning approach to predict the artists of paintings. Automatic analysis has seen application in the archeology domain too. Work done by Cintas et al. [3] uses a dataset about Iberian ceramics and they demonstrate the efficiency of the automatic classification of pottery vessels. In [31] the authors combine image and semantic embeddings for classification of statues. Classification tasks include the style, type, dimension, century, and material of the statues. Deep learning techniques have been applied to recognize characters in images of art history [18]. Their dataset consists 2,787 images of artworks of specific iconography and their transfer learning strategy with deep CNN models outperforms the traditional ML techniques for their character recognition task.

Earlier work has shown great interest of crowdsourcing applications to cultural heritage domain. Oosterman et al [25] focus on specific artwork annotation that requires domain knowledge, comparing annotations of crowd workers to experts. The task of annotating collection of prints depicting flowers from museums showed that there is a clear relation between difficulty of the annotation task with the performance of the crowd workers. Knowledge intensive tasks require employing trained crowd workers. "Accurator" [10] is a nichesourcing methodology that proposes to tailor the annotation tools to a domain and to address specific crowd communities. This methodology has shown to collect high quality of annotations for a variety of domains. Several other works [28, 36] focus on image tagging of artworks. The conducted experiment by Traub et al [36] to annotate oil paintings show that introducing gamification and simplifying an expert task into non-expert task can enable ordinary crowd workers to accomplish nearly what experts can.

5 CONCLUSIONS AND FUTURE WORK

In this work, we have addressed the task of categorizing images from cultural heritage collections. We presented our hybrid human-machine framework for image categorization. This method aggregates the predictions of multi-input deep learning model and the inferred true annotations from multiple crowd users. The deep learning model uses visual features extracted from the images and features extracted from text in the image metadata. We found that our method of adding Wikipedia classes of the concepts extracted in the text improves the classification accuracy. Moreover, incorporating annotations from crowd users and applying a weighted aggregation additionally improved the results. In summary, our results have confirmed the assumption that the hybrid aggregation method is an effective approach to combining machine learning with crowd annotation skills. This method is helpful for organizations like GLAMs that maintain data repositories in the cultural heritage domain, and which have many active participants. Currently, the NotreHistoire platform has thousands of registered users, and several hundred active participants. Hence, the proposed strategy is well applicable to categorize the images in such a context.

The size of the dataset is a possible limitation that especially affects the deep learning approach. In principle, deep neural network models require more data, therefore, our future work will focus on increasing the size of the dataset to improve the accuracy. We identified that the collection metadata has additional issues such as missing high quality *image tags* and the *period* which defines the temporal decade of the images. We plan to deploy our method to also address these additional missing data.

ACKNOWLEDGMENT

This work was partly funded by the Hasler Foundation in the context of the project *City-Stories*. We would like to thank the NotreHistoire project and FONSART for delivering a data testbed.

REFERENCES

[1] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. 2013. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing* 17, 2 (2013), 76–81.

[2] Abdelhak Belhi, Abdelaziz Bouras, and Sebti Foufou. 2018. Leveraging known data for missing label prediction in cultural heritage context. *App. Sciences* (2018).

[3] Celia Cintas, Manuel Lucena, José Manuel Fuertes, Claudio Delrieux, Pablo Navarro, Rolando González-José, and Manuel Molinos. 2020. Automatic feature extraction and classification of Iberian ceramics based on deep convolutional networks. *Journal of Cultural Heritage* (2020), 106 – 112.

[4] ICOMOS International Cultural Tourism Committee et al. 2002. International Cultural Tourism Charter: Principles and Guidelines for Managing Tourism at Places of Cultural and Heritage Significance. 13 June 2013.

[5] Joe Cox, Eun Young Oh, Brooke Simmons, Gary Graham, Anita Greenhill, Chris Lintott, Karen Masters, and James Woodcock. 2015. Doing good online: An investigation into the characteristics and motivations of digital volunteers. *Leeds University Business School Working Paper* 16-08 (2015).

[6] Maria Daltayanni, Luca de Alfaro, and Panagiotis Papadimitriou. 2015. Workerrank: Using employer implicit judgements to infer worker reputation. In *Proc. of the* 8^{th} *ACM International Conference on Web Search and Data Mining*. 263–272.

[7] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 1 (1979), 20–28.

[8] Gianluca Demartini. 2015. Hybrid human–machine information systems: Challenges and opportunities. *Computer Networks* 90 (2015), 5–13.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE CVPR*. 248–255.

[10] Chris Dijkshoorn, Victor De Boer, Lora Aroyo, and Guus Schreiber. 2017. Accurator: nichesourcing for cultural heritage. (2017).

[11] PN Druzhkov and VD Kustikova. 2016. A survey of deep learning methods and software tools for image classification and object detection. *Pattern Recognition and Image Analysis* 26, 1 (2016), 9–15.

[12] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. 2013. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*. 1–4.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2015. Region-based convolutional networks for accurate object detection and segmentation. *IEEE tx on pattern analysis and machine intelligence* 38, 1 (2015), 142–158.

[14] Eric Hand. 2010. People power: networks of human minds are taking citizen science to a new level. *Nature* 466, 7307 (2010), 685–688.

[15] Rose Holley. 2010. Crowdsourcing: How and why should libraries do it? *D-Lib magazine* 16, 3/4 Ma (2010).

[16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. (2017).

[17] Jose Llamas, Pedro M Lerones, Roberto Medina, Eduardo Zalama, and Jaime Gómez-García-Bermejo. 2017. Classification of architectural heritage images using deep learning techniques. *Applied Sciences* 7, 10 (2017), 992.

[18] Prathmesh Madhu, Ronak Kosti, Lara Mührenberg, Peter Bell, Andreas Maier, and Vincent Christlein. 2019. Recognizing Characters in Art History Using Deep Learning *(SUMAC '19)*.

[19] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

[20] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*. 1–8.

[21] Archana Nottamkandath, Jasper Oosterman, Davide Ceolin, Wan Fokkink, et al. 2014. Automated Evaluation of Crowdsourced Annotations in the Cultural Heritage Domain.. In *URSW*. 25–36.

[22] David Oleson, Alexander Sorokin, Greg Laughlin, Vaughn Hester, John Le, and Lukas Biewald. 2011. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Workshops at the* 25^{th} *AAAI Conf. on Artificial Intelligence*.

[23] Alex C Olivieri, Roland Schegg, and Maria Sokhn. 2016. Cityzen: a social platform for cultural heritage focused tourism. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*. 129–136.

[24] Johan Oomen and Lora Aroyo. 2011. Crowdsourcing in the cultural heritage domain: opportunities and challenges. In *Proceedings of the 5th International Conference on Communities and Technologies*. 138–149.

[25] Jasper Oosterman, Archana Nottamkandath, Chris Dijkshoorn, Alessandro Bozzon, Geert-Jan Houben, and Lora Aroyo. 2014. Crowdsourcing Knowledge-Intensive Tasks in Cultural Heritage. In *ACM Web Science Conference*.

[26] Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra. 2017. Multi-label music genre classification from audio, text, and images using deep features. (2017).

[27] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[28] Dimitris Paraschakis and Marie Gustafsson Friberger. 2014. Playful crowdsourcing of archival metadata through social networks. (2014).

[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[30] Kostas Proedrou, Ilia Nouretdinov, Volodya Vovk, and Alex Gammerman. 2002. Transductive confidence machines for pattern recognition. In *European Conference on Machine Learning*. Springer, 381–390.

[31] Benjamin Renoust, Matheus Oliveira Franca, Jacob Chan, Noa Garcia, Van Le, Ayaka Uesaka, Yuta Nakashima, Hajime Nagahara, Jueren Wang, and Yutaka Fujioka. 2019. Historical and Modern Features for Buddha Statue Classification. In *SUMAC '19*. 23–30.

[32] Shaban Shabani and Maria Sokhn. 2018. Hybrid machine-crowd approach for fake news detection. In *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 299–306.

[33] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

[35] Jennifer Trant. 2009. Tagging, folksonomy and art museums: Early experiments and ongoing research. (2009).

[36] Myriam C Traub, Jacco van Ossenbruggen, Jiyin He, and Lynda Hardman. 2014. Measuring the effectiveness of gamesourcing expert oil painting annotations. In *European Conference on Information Retrieval*. Springer, 112–123.

[37] G Vinodhini and RM Chandrasekaran. 2012. Sentiment analysis and opinion mining: a survey. *International Journal* 2, 6 (2012), 282–292.

[38] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. recaptcha: Human-based character recognition via web security measures. *Science* 321, 5895 (2008), 1465–1468.