Human-Machine Collaboration for Face Recognition

by

Saurabh Ravindranath, Rahul Baburaj, Vineeth Balasubramanian, Nageswara Rao Namburu, Sujit Prakash Gujar, C V Jawahar

in

7th ACM India Joint International Conference on Data Science Management of Data, COMAD/CODS, ACM, 2020

Report No: IIIT/TR/2020/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
January 2020

Human-Machine Collaboration for Face Recognition

Saurabh Ravindranath
University of Southern California
United States of America
saurabhr@usc.edu

Rahul Baburaj
IIT, Madras
Chennai, India
rahuledachali@gmail.com

Vineeth N. Balasubramanian
IIT Hyderabad
Hyderabad, India
vineethnb@iith.ac.in

NageswaraRao Namburu
IIIT Hyderabad
Gachibowli, Hyderabad, India
nageswara.na@research.iiit.ac.in

Sujit Gujar
IIIT Hyderabad
Gachibowli, Hyderabad, India
sujit.gujar@iiit.ac.in

C. V. Jawahar
IIIT Hyderabad
Gachibowli, Hyderabad, India
jawahar@iiit.ac.in

Figure 1: Our system models human experts realistically, that is, having varying accuracies. We show that this modeling allows our system to obtain better accuracies at a lower cost in a real-world setting, when compared to systems making machine-only predictions (Method 1) or systems modeling human accuracies homogeneously (Method 2), similar to [24].

ABSTRACT

Despite advances in deep learning and facial recognition techniques, the problem of fault-intolerant facial recognition remains challenging. With the current state of progress in the field of automatic face recognition and the in-feasibility of fully manual recognition, the situation calls for human-machine collaborative methods. We design a system that uses machine predictions for a given face to generate queries that are answered by human experts to provide the system with the information required to predict the identity of the face correctly. We use a Markov Decision Process for which we devise an appropriate query structure and a reward structure to generate these queries in a budget or accuracy-constrained setting. Finally, as we do not know the capabilities of the human experts involved, we model each human as a bandit and adopt a multi-armed bandit approach with consensus queries to efficiently estimate their individual accuracies, enabling us to maximize the accuracy of our system. Through careful analysis and experimentation on real-world data-sets using humans, we show that our system outperforms methods that exploit only machine intelligence, simultaneously being highly cost-efficient as compared to fully manual methods. In summary, our system uses human-machine collaboration for face recognition problem more intelligently and efficiently.

KEYWORDS

facial recognition techniques, Markov Decision Process, crowdsourcing, multi-armed bandits

ACM Reference Format:
Saurabh Ravindranath, Rahul Baburaj, Vineeth N. Balasubramanian, NageswaraRao Namburu, Sujit Gujar, and C. V. Jawahar. 2020. Human-Machine Collaboration for Face Recognition. In *7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020), January 5–7, 2020, Hyderabad, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3371158.3371160

CoDS COMAD 2020, January 5–7, 2020, Hyderabad, India

ACM ISBN 978-1-4503-7738-6/20/01…$15.00
https://doi.org/10.1145/3371158.3371160

1 INTRODUCTION

Face recognition by machines has improved considerably in the past few years, highlighted by the ability to recognize frontal faces with very high accuracy in still images taken in controlled environments [26, 29]. However, these systems have been shown to be sensitive to different factors, such as viewpoint variations, which human recognition is largely robust to. Further, machine learning-based systems are prone to adversarial attacks and spoofs, which pose a major security threat in many applications. The friction between accuracy and interpretability of deep learning models pertains to

the trade-off between being able to accomplish complex tasks and understanding how these tasks are accomplished. The most intuitive way to tackle these uncertainties is to include humans in the loop.

The proposed human-in-the-loop approach seeks to combine the enormous data-crunching capability of machines with the domain expertise of humans. Humans contribute by providing knowledge and capabilities that are difficult to model for a machine learning system. Simultaneously, a machine would be able to handle enormous data-sets that are far beyond the reach of expert observation. Typical examples are biometric security and classification problems [19, 33, 34], where the task is too ambiguous for a purely mechanical solution and too large for even a large team of human experts.

For sensitive security applications such as face recognition from security cameras [22], appropriate human involvement in *run time* can be a critical necessity. As the number of identities grows, it becomes increasingly important to find a structured way to include humans-in-the-loop.

In this work, we develop a structured framework to appropriately combine the prediction capabilities of a machine learning model with that of a diverse group of humans while optimizing over the accuracy-cost trade-off. To this end, we propose a *Markov decision process (*MDP*)* [6, 9, 16, 17, 27] that takes the probability predictions from the machine learning model as input and use humans to refine its predictions through specific queries. To involve a diverse group of human experts, we have to consider the variation in the quality of the individual experts. That is, selecting different human experts for the same task may result in variation in the accuracy. The task of selecting human experts needs to consider the trade-off between exploration and exploitation. Existing techniques do not deal with this real-world problem of heterogeneous human expert qualities. Similar to many previous methods that employ human-in-the-loop methods , we use a *multi-armed bandit (*MAB*)* [3, 5, 15, 30, 31] mechanism to select the human best suited for each task. As represented in Figure 1 the realistic modeling of human experts helps our model to increase accuracy while decreasing the overall cost. Our system outperforms the systems that rely only on machine predictions and also the methods which model humans as having homogeneous accuracies. Overall, the major contributions of our work are as follows:

- We formulate an MDP to model machine and human knowledge in a single framework.
- We incorporate human accuracies, estimated on the go using MAB and consensus agreement approaches, to choose the best human for a particular task.
- We propose a novel combination of algorithms for tackling the problem of face recognition that combines human and machine intelligence ingeniously. We employ an MDP framework for query optimization with an MAB for the quality estimation of human experts.

To obtain the accuracy estimates of each human involved, we first make "gold queries," queries for which the correct answer is already known. However, as these gold queries are expensive, we use these only to obtain a rough accuracy estimate and then fine-tune these estimates using certain consensus agreement queries. Our results demonstrate that we perform better than both machine learning based models and the human workforce at tackling face recognition. Moreover, although here we demonstrate our method on the face recognition problem, it can be easily generalized to other applications such as object detection, image classification, and various segmentation tasks.

An interesting result of using MDP in the engagement of humans in our system is the increased flexibility it offers in terms of the level of involvement. The ability to devise appropriate reward functions to optimize particular problems opens up a new dimension of issues that can be solved, relating to the systematic deployment of Artificial Intelligence components in manual scenarios. As observed in our experiments, increasing the budget of the system increases human involvement and the final accuracy in a structured manner. Such a property would greatly alleviate issues that arise when Artificial Intelligence systems replace manual alternatives.

Coming to the issue of privacy, we assumed that all are publicly known people, so there will not be an issue. It is outside the scope of this work to consider the privacy of unknown people.

Before we move onto details, it should be noted that to best of our knowledge, human-in-the-loop approaches are widely used either in collecting labeled data or in the active learning scenario where the machine learning model is updated with the involvement of human supervision during *training* [8, 35]. However, in this work, we are proposing to use humans when a trained machine learning algorithm is put in a system and at *run-time*, has low confidence in the answer. This is very much necessary in security applications such as face recognition. We believe the systematic study, which we do in this paper, has many more applications, e.g., machine translation systems.

2 RELATED WORK

Different works have been carried out in the field of face recognition using deep learning techniques, some of which produce robust results in few-shot learning and other challenging scenarios. However, these methods are not fully robust to viewpoint variations, which is a fundamental problem with CNN architectures. Although capsule network architectures [11, 25] are viewpoint invariant, their scalability is still a question. There is a vast corpus of work in face verification and recognition. Reviewing it is out of the scope of this paper, so we briefly discuss the most relevant recent work. The works of [28, 29, 37] employ a complex system of multiple stages which combines the output of a deep convolutional network with PCA for dimensionality reduction and an SVM for classification. Taigman *et al.* [29] explicitly models a 3D face and derives a face representation from a deep neural network. Schroff *et al.* [26], on the other hand, uses a deep convolutional network trained to optimize the embedding space itself directly. The learned representation space in all these can be used for face recognition, employing a classifier in the end.

In recent years, the computer vision community has seen a burst of interest in deep learning based solutions to image classification problems. Different active learning approaches are also used to find the most diverse data-set effectively. In the above cases, human involvement is present only in the annotation phase. Unlike these methods, our work focuses on using human input in the testing phase. Works most similar to our approach are those who combine human help and

computer vision for tasks such as fine-grained image classification [4, 33], object detection [24], and image segmentation. Works such as [7, 32] model a human-computer interactive system and optimizes human cost versus annotation accuracy. However, these methods either do not incorporate individual human expert accuracy into the formulation or do not include human in the annotation phase.

Crowd-sourcing platforms offer an inexpensive method to capture human knowledge and understanding for a vast number of visual perception tasks. Selecting optimal works is challenging problem and there is lot of research happening in this direction, e.g., [10, 14, 18, 21, 23, 31] and references cited there-in. When qualities of the crowd workers (human experts) are unknown, the most natural tool is Multi-Armed Bandits (MAB). MAB [3] based work such as [12, 15, 30, 31] are used in this scenario, where an optimal subset of experts is selected for the task such that the outcome after aggregating their opinions guarantees a target level of accuracy. We employ MAB and consensus agreement algorithms used in these works to learn the quality of predictions made by each human expert.

The human-machine collaboration framework for object annotation proposed by Russakovsky *et al.*[24] is closely connected to our approach. The input to their system is an image, and a set of annotation constraints, and the output is a set of object annotations informed by computer vision and human feedback. We employ their idea of using an MDP for effectively incorporating human and machine knowledge and adapt it to the problem of face recognition. Our method hence has a similar state space, a more trivial action space, but novel reward functions and look-ahead criteria for the different dimensions we optimize on. We draw upon all the above works to integrate machine learning models, human feedback, and their quality estimation in the most effective way.

3 PROBLEM FORMULATION

We demonstrate a policy that maximizes the accuracy of face recognition using human expert input at test time, in addition to the machine predictions. We model human experts in this system by assuming an accuracy value for each expert and fine-tuning them as feedback is collected. Our system consists of the following components:

BM - Machine classifier: (Baseline Model - Computer vision face recognition model): This is the face recognition model which provides us with the initial set of predictions. The input to this system is an image of a face, and the output is a vector of probabilities of the face belonging to the respective identities.

HIL-MDP - Efficient query selection: (human in the loop - MDP for query selection) This component decides the queries to ask the human expert to maximize accuracy given a cost constraint in time or minimize the human expert time required while guaranteeing a given accuracy. We use a Markov Decision Process to achieve this.

HIL-MAB - Annotator quality estimation and selection: (human in the loop - multi-armed bandit for query addressing) This component deals with selecting the human expert best suited for the given task. This is carried out by estimating and fine-tuning the accuracy estimates of the different human experts by assigning some gold and consensus queries to them and obtaining their feedback. A Multi-armed Bandit approach is used to select the human expert whose accuracy values are to be fine-tuned for each query.

4 OUR SYSTEM FOR FACE RECOGNITION

In this section, we explain how we tackle a face recognition problem by combining machine intelligence with human intelligence. The target image flows in our system until the face is identified as follows. The image of a face is given as input to the BM component. This component makes predictions and gives the confidence measures of the different identities being the same as the one in the image. The HIL-MDP component uses these predictions to select the queries to be asked, such that the final accuracy is maximized, and the final cost is minimized. Finally, the HIL-MAB component assigns these queries to the different human experts based on their qualities, which is inferred through the feedback it obtains from them. The diagrammatic representation can be observed in Figure 2. We now explain how these three modules work in our system.

4.1 BM

We use the OpenFace face recognition [1] methodology to provide initial predictions for each face for our system. For each image, we perform a face alignment. The FaceNet [26] pre-trained embedding space is then employed to represent the face on a 128-dimensional unit hyper-sphere. This embedding space is an end-to-end learned system that employs a triplet loss that directly reflects what we wish to achieve in face recognition. A Support vector machine (SVM) classifier trained on this set of images is then used to recognize the faces. This system outputs the list of predictions and their confidence values, which are used as initial predictions for our setup.

4.2 HIL-MDP

Notation	Meaning
$I_1, I_2, ..., I_n$	The set of all possible identities
$I_l \vert s$	The identity with the highest probability of being the face in the image, at state s
I_c	The correct identity of the face in the image
$P(I_k \vert s)$	Probability that the k^{th} identity is the identity of the face in the image at state s
$P(TP)$	Probability that the human expert will pick the correct identity if it is present in the query.
$P(TN)$	Probability that the human expert correctly identifies that none of the queried identities corresponds to the input face.

Table 1: List of symbols used in describing the working of the HIL-MDP component and what they mean.

Multiple types of queries can be assigned to the human experts to obtain information about the identity of the face in the image. These queries would have varying costs, depending on their complexity. In our system, we define these costs as the amount of human expert time required to complete them.

Here, the queries we ask are of the type: Given k Identities, and the image of the face, determine if the face belongs to any of these

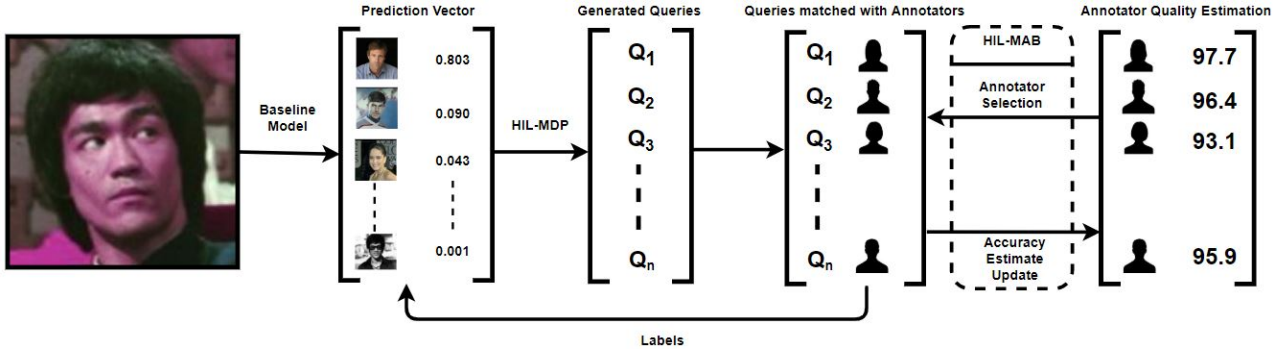

Figure 2: System overview: The inputs, outputs and interaction between the different components can be seen. BM **makes predictions based on the image,** HIL-MDP **selects the queries to be asked and** HIL-MAB **chooses the human experts to ask these queries to, based on the feedback they give.**

identities and if yes, which identity. Many such queries would have to be generated for each image, in case of different human expert input. These queries would only depend on the current confidence scores of the different identities being the identity in the image. Further, for each value of k, there would be a different cost associated with the query, as more human expert time would be required to answer some queries than others. The aim of the system would be to minimize the total cost of the queries while choosing the queries that could possibly give the most information to the system. Due to these reasons, we use a Markov decision process (MDP) to generate these queries. This approach is similar to that found in Russakovsky *et al.* [24]. However, we modify the state and action spaces to that of the face recognition problem. We also have created different reward functions to optimize for different sub-problems, them being:

- Obtaining the best possible estimated accuracy of prediction, given a fixed budget
- Expending the least human effort to achieve a given estimated accuracy constraint

Since such an MDP would not have a fixed end-state, we have set a horizon or look-ahead as the number of actions it would search through, from the given start state. This would effectively set a limit on the number of queries that can be posed about a given image. An MDP is defined as a tuple of the following form (S, A, T, R), the sets being:

State Space, S. : The state-space of the MDP is the probability vector of the identity predictions for the given image. The prediction vector output of the computer vision model will be the start state.

Action Space, A. : The actions of the MDP are the set of queries we will be assigned to the human experts. We have a set of k possible actions, $a_1, a_2, ..., a_k$ from every state, defined as follows:

a_k: Given k identities, $I_1, I_2, ..., I_k$, and the image, determine if the face belongs to any identity and if yes, which identity. The cost of this action would be a monotonically increasing function of k, i.e., $c(k)$.

As seen in Figure 3, the action a_k will have $k+1$ possible answers, leading to $k+1$ different identity prediction vectors. However, k of these answers is the human expert claiming that the identity of the input face is one of the identities present in the query, thus determining the final answer. This implies that no more queries would have to be asked on the image. Hence, the action a_k would lead to k terminal state and 1 non-terminal state, that is when the human expert answers that none of the identities in the query is the identity of the face in the image.

Transition probabilities, T. : The state reached on asking a query, a_k to a human expert would depend on the probability of the human expert picking the correct answer, and the correct identity being present in the query. Thus, the probability to transition to different states when query a_k is posed are as follows:

- I_c is present in $I_1, I_2, ..., I_k$, and the human expert labels correctly: In this case, with a probability of

$$P(TP)\sum_{j=1}^{k} P(I_j|s) \tag{1}$$

  the state will transition to: $P(I_m|s') = 1$, where $I_m = I_c$, and $P(I_j|s') = 0, \forall j \neq m$. This implies that the system produces the final answer, I_c
- I_c is not present in $I_1, I_2, ..., I_k$, and the human expert labels correctly: In this case, with a probability of

$$P(TN)(1 - \sum_{j=1}^{k} P(I_j|s)) \tag{2}$$

  the state will transition to: $P(I_m|s') = 0, \forall 1 \leq m \leq k$, and $\sum_{j=1}^{k} P(I_j|s')$, $\forall j \neq m$ will be added to $P(I_j|s')$, $\forall j \neq m$ in a weighted manner

Reward Function, R. : The reward function used would depend on the quantity being optimized. In our work, we have created two separate sub-problems, each optimizing a different quantity, that is, either human expert cost or estimated accuracy of the identity.

(1) **Given a fixed budget, to maximize the estimated accuracy:** Estimated accuracy of a state s, on performing an action a and moving to states s' stochastically, can be calculated as follows:

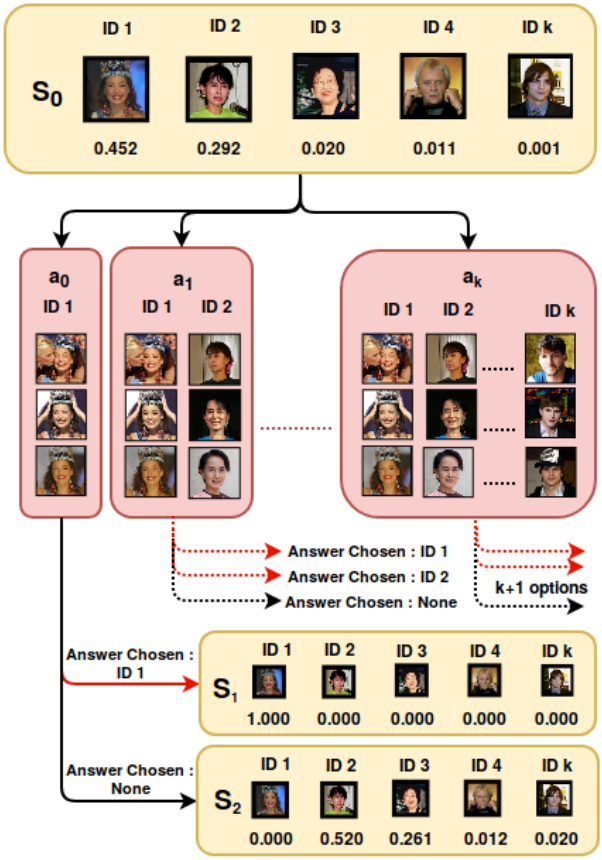

Figure 3: MDP state transitions: State S_0 would be the prediction vector output by the classifier, where ID 1, . . . , ID k are the different identities. Action a_1 would be posed to the human expert by giving 3 images of ID 1, to query whether ID 1 is the identity present in the input image or not. All the human expert answers that lead to terminal states are marked with red arrows.

$$V(s,a) = \sum_{s'} P(s'|s,a)(R(s') + \max_{a'} V(s',a')) \tag{3}$$

which, in this case would be:

$$V(s,a) = [P(TP)\sum_{j=1}^{k} P(I_j|s)] \times 1 + P(TN)[1 - \sum_{j=1}^{k} P(I_j|s)] \times (P(I_l|s') + \max_{a'} V(s',a')) \tag{4}$$

where a' would be an action from state s'.
Thus, here, the reward function, $R(s)$ will be $P(I_l|s)$.
To not exceed the given cost constraint, we limit the horizon of the MDP such that actions, which on inclusion exceed the budget, are not considered in calculating the estimated accuracy.

(2) **Expending the least human effort to achieve a given accuracy constraint:** As our goal is to minimize the cost of the MDP given an accuracy constraint, we have set the reward function, $R(s,a)$ to simply as $R(s,a) = -cost(a)$.
To stay within the given estimated accuracy constraint, $A_0 >= V(s_0)$, we modify Equation (4) to obtain the following expression:

$$A_1 >= \frac{A_0 - P(TP)\sum_{j=1}^{k} P(I_j|s) \times 1}{P(TN)(1 - \sum_{j=1}^{k} P(I_j|s))} \tag{5}$$

Here, A_i would be the estimated accuracy constraint to be satisfied in the i^{th} look-ahead of the MDP.
Having the values of A_i would allow us to eliminate all action sequences which do not meet the given accuracy constraint.

4.3 HIL-MAB

Here, the task is to find the best human for each query by estimating the accuracy of each human expert. To this end, each human can be modeled as a bandit, with an unknown accuracy.

As the accuracy of each human expert is unknown in the beginning, we first obtain a rough estimate by posing certain gold queries, i.e. face images where the identity is known. To fine-tune these accuracies, we then pose certain queries to a subset of human experts to arrive at a consensus. Here, the consensus answer can be assumed to be the true answer as a high accuracy level is being assured. Such an assumption is similar to the assumptions made in [15, 20]. Due to this, the queries are split into three types, and the algorithm to select the human experts for each type of query is as follows:

Notation	Meaning
na	Total number of human experts
ca_i	Number of queries answered correctly by human expert i
ta_i	Number of queries answered by human expert i
q_i	The quality of the human expert, calculated as $q_i = \frac{ca_i}{ta_i}$

Table 2: List of symbols used in describing the working of the HIL-MAB component and what they mean

Type A - Gold queries. Gold queries are expensive, and thus, we aim to reduce the number of gold queries our system uses. This is done by first passing the gold queries through the MDP to generate queries and then assigning each query to all the experts present, rather than a subset of them. As we know, the correct answers, ca_i, and ta_i can be updated, and thus, for each query, we can update the q_i values of all the human experts. In our work, we set a certain fraction of the queries as gold queries; however, future work can be done to optimize the number of gold queries, similar to the work done by Chien-Ju Ho *et al.* [12].

Type B - Consensus queries. Estimating the accuracy with high precision would require numerous gold queries. To avoid this, we aim to obtain "consensus labels" that would be more accurate than the label given by a single human expert.

We only perform these accuracy updates on queries of MDP type a_1, where only one identity is given to the human expert, and he/she is asked to identify if this identity is the same as the one in the input face image or not. This is done because, when such a query is asked an odd number of human experts, a consensus is guaranteed to emerge, as there are only two possible options. Once a consensus is reached, we assume the consensus is the correct answer, and accordingly, the ca_i, ta_i, and thus the q_i values can be updated for the human experts the query was directed to.

To maintain the balance between exploring the accuracies of the lesser assigned human experts and exploiting the high accuracies of more assigned human experts, we use a multi-armed bandit solution, namely the upper confidence bound algorithm [2, 5]. The algorithm greedily picks the human expert i with the highest UCB value. The UCB value for a human expert i, UCB_i is computed as follows:

$$UCB_i = q_i + \sqrt{\frac{2\log(\sum_{i=1}^{na} ta_i)}{ta_i}} \tag{6}$$

The confidence bound increases with the total number of times we pick any human expert but decreases with the number of times we have picked a particular human expert. This ensures that although the number of times each human expert is picked tends to infinity as the total number of trials tends to infinity, the balance still exists between exploration and exploitation. Thus, for queries of this type, three human experts having the highest UCB_i values are selected.

Type C - Non-update queries. The accuracy values estimated using the previous two types of queries are used to select the human experts used in these types of queries. A single human expert with the highest current q_i value is selected to answer the query. The responses to these queries are not used to update the q_i values of human experts.

To summarize, the procedure to select appropriate human experts for specific queries can be outlined as follows:

Algorithm 1 Query Distribution

1: **if** Type A Query **then** Assign to all human experts
2: **else if** Type B Query **then**
3: Compute the UCB value of each human expert using Equation 6
4: Assign query to 3 human experts having highest UCB values
5: **else if** Type C Query **then**
6: Assign query to human expert with highest q_i value

5 EXPERIMENTS

In this section, the experimental results of the above system can be observed.

5.1 data-sets used

LFW data-set. We use a subset of the "labeled faces in the wild" data-set [13], training our baseline model on 85 identities, with a training set of size 2,285 images and testing size of 1,315 images. Our classifier achieves an accuracy of 88.6% on the test set using a facenet embedding space and SVM classifier face recognition system.

CASIA WebFace data-set. We use a subset of the "CASIA WebFace data-set" data-set [36], training our baseline model on 500 identities, with a training set of size 15,000 images and testing size of 5,000 images. Our classifier achieves an accuracy of 84.32% on the test set.

5.2 Simulated Experiment Details

Action cost determination. In our system, we require costs to be assigned to the different possible queries chosen by the MDP. Each MDP query would be of the form a_k, where the human expert would have to determine whether one of the identities is the same as that in the given image, or not, if yes, to pick the identity. In this context, we define cost to be the amount of human expert time spent on the action. Note that, the monetary rewards will vary from the demographics of the human experts and hence, we work with time taken by the human experts as costs. In our experiments, we will not be accounting for idle human expert time, and thus the total cost will be the summation of the time spent on all the queries. To determine these costs, we assigned 6,500 queries to 8 different human experts and on averaging the results, the following are the costs obtained:

Action	a_1	a_2	a_3	a_4	a_5
Cost (in seconds)	3.8	5.7	8.9	11.3	14.7

Table 3: Action Cost Ground-Truth: The average time taken by the humans to answer each HIL-MDP query.

Human Annotator Simulation. To test our system on larger data-sets, and to run it on different sets of constraints, we built a system that simulates human experts. In this simulation, each human expert is assigned an accuracy, $a : 0 \leq a \leq 1$, and with a probability of a, the right answer is returned, and with a probability of $1 - a$, a random wrong answer is returned.

To ensure such simulations give accurate results, we estimated the accuracy of 8 human experts, over 476 queries each, and their accuracies were: {94.53%, 91.19%, 95.80%, 95.38%, 98.95%, 94.53%, 97.27%, 97.05%}. These are the accuracy values of the human experts in all the simulations that follow.

5.3 Load Balancing

In our system, type A and type B queries have a mechanism of being distributed among the different human experts (Type A being assigned to all human experts, and type B being assigned using the UCB algorithm), whereas type C queries do not. This results in our system using multiple human experts inefficiently. To alleviate this issue, we introduce a "balance factor", b_f into our system. This factor ensures that the difference in the number of type C queries between the human experts having the i^{th} and the $(i+1)^{th}$ highest type C queries never exceed b_f. This ensures that queries of type C are also distributed in a balanced manner among human experts. As b_f increases, the number of queries assigned to each human expert gets skewed, and the accuracy of the system increases. The results when simulations are run without budget or accuracy constraints can be observed in Table 4.

b_f	Accuracy	Total Cost	Range	SD (σ)
LFW data-set				
50	96.75%	11524.90 sec	575	199.42
100	97.32%	8342.70 sec	952	334.72
200	98.11%	8293.49 sec	1112	373.68
No b_f	98.56%	15954.50 sec	4372	1431.20
CASIA WebFace data-set				
50	95.5%	55452.10 sec	377	120.18
200	95.84%	55750.60 sec	1578	504.87
500	96.33%	65681.90 sec	5338	1928.79
No b_f	97.29%	73076.40 sec	18944	6214.09

Table 4: Effect of balance factor on accuracy, cost and query distribution among human experts, measured using range and standard deviation (SD) of the number of queries assigned to the human experts.

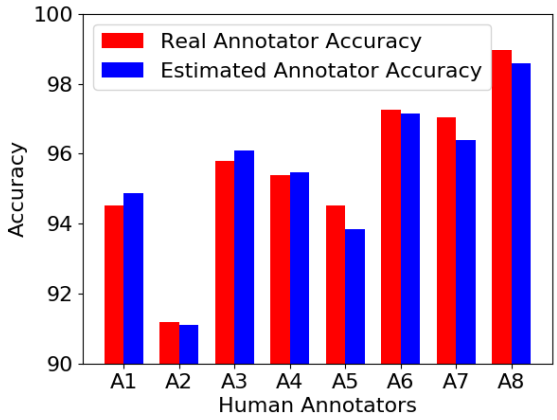

Figure 4: Graph showing how our system can accurately predict accuracies of simulated humans. Here, A1, A2, ..., A8 are the human experts whose accuracy measure has been estimated.

5.4 Impact of HIL-MAB Component

To observe the impact of HIL-MAB, we run the simulation without budget constraints and load balancing, and with and without the HIL-MAB component. The results can be observed in Table 5. On the addition of the HIL-MAB component, the difference between the qualities of the different human experts is capitalized, thus assigning queries accordingly, resulting in higher accuracies. It can be observed that the total cost required also decreases, as when human experts with higher accuracies are picked to answer queries, more information is gained, and thus, the HIL-MDP would have to make fewer queries to improve the prediction.

Experiment	Accuracy		Total Cost (sec)	
	without MAB	with MAB	without MAB	with MAB
LFW	95.05%	98.56%	16996	15954
CASIA	93.41%	97.29%	76263	73076

Table 5: Results obtained with and without the HIL-MAB component: There is significant in the accuracies obtained when using the HIL-MAB component.

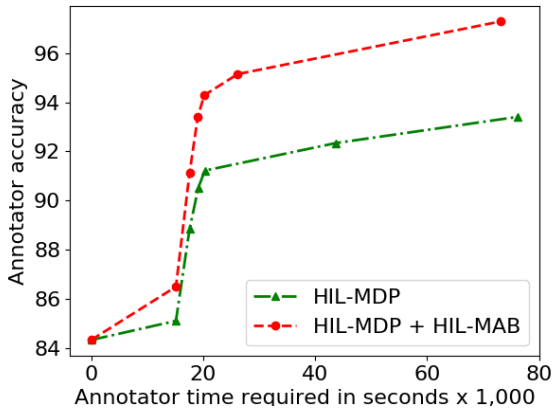

Figure 5: Accuracy vs Cost graph for different budget constraints simulated on the CASIA WebFace data-set. The points, from left to right are budget constraints of 0, 4, 8, 12, 16 and Unlimited human expert seconds per image respectively.

5.5 Impact of HIL-MDP Optimization

On the CASIA WebFace data-set, we ran simulations with different budget constraint values while including and excluding HIL-MAB, as seen in Figure 5. As we use HIL-MDP to determine the queries to be asked the humans, we can see how increasing the available budget causes the system to either increase the number of queries asked or ask more expensive queries, thus increasing the accuracy obtained. It can also be seen how the addition of the HIL-MAB component consistently provides better results.

5.6 Using Real Human Input

We ran the system with eight real human experts on the LFW data-set. We did not assign a budget constraint. A balance factor of $b_f = 50$ was set to minimize human expert wait time. Through gold queries and consensus queries, the HIL-MAB component estimated the accuracy measures of the human experts as {94.53%, 91.19%, 95.80%, 95.38%, 98.95%, 94.53%, 97.27%, 97.05%}. Out of the 1315 queries, 80 of them were gold queries.

The final accuracy achieved under these conditions was 95.16%, which follows the trends that we have seen in our simulations.

6 DISCUSSION AND CONCLUSIONS

We presented a novel system that combines the ability of machine learning approaches to handle large volumes of data with superior human accuracy in the domain of facial recognition. Using real human experts, we show that our system can obtain an accuracy of

Figure 6: Impact of human input: We can see how although the classifier fails to give the correct input, the correct answer is present in the top-k results and is easily identifiable by a human expert.

95.16% on a subset of the LFW data-set, whereas the machine-only approach obtained an accuracy of 88.6%.

We show that we can introduce different degrees of load balancing in our system, which allows us to more efficiently use the multiple human experts present at the cost of accuracy.

We highlight the importance of the MAB component to assign the human expert to be queried. Through our experiments, it is seen how accurately estimating the accuracy of each human expert and assigning queries appropriately increases the overall accuracy of our system and also decreases the cost incurred. In the future, we will go for contextual multi-armed bandits that consider time-varying as well as demographic-based accuracies of human experts.

One of the results of using an MDP in our system is that there is scope for obtaining optimal results under varying constraints. As seen in our experiments, different sets of queries are generated according to the specified budget, which correspondingly increases or decreases the final accuracy of the system.

We believe that such a combination of the machine input, along with MDP and MAB methods to obtain human input, can be applied to other machine learning tasks.

In this work, we treated workers having a fixed accuracy. However, further MAB theory needs to be developed as though tasks are homogeneous can greatly vary in difficulty. We leave it for future work, how to efficiently do explore-exploit trade-off in such settings.

The proposed approach is very generic and could be used in plenty of applications such as machine translation.

REFERENCES

[1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications*. Technical Report. CMU-CS-16-118, CMU School of Computer Science.

[2] Peter Auer. 2003. Using Confidence Bounds for Exploitation-exploration Tradeoffs. *J. Mach. Learn. Res.* 3, Nov (2003), 397–422.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2-3 (2002), 235–256.

[4] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. 2010. Visual recognition with humans in the loop. In *European Conference on Computer Vision*. 438–451.

[5] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5, 1 (2012), 1–122.

[6] Peng Dai, Daniel Sabey Weld, et al. 2010. Decision-theoretic control of crowdsourced workflows. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

[7] Suyog Dutt Jain and Kristen Grauman. 2013. Predicting sufficient annotation strength for interactive foreground segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1313–1320.

[8] Meng Fang, Jie Yin, and Dacheng Tao. 2014. Active learning for crowdsourcing using knowledge transfer. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

[9] Ben Goodrich and Itamar Arel. 2012. Reinforcement learning based visual attention with application to face detection. (2012), 19–24.

[10] Sujit Gujar and Boi Faltings. 2017. Auction Based Mechanisms for Dynamic Task Assignments in Expert Crowdsourcing. In *Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets*, Sofia Ceppi, Esther David, Chen Hajaj, Valentin Robu, and Ioannis A. Vetsikas (Eds.). Springer International Publishing, Cham, 50–65.

[11] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. 2011. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*. Springer, 44–51.

[12] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. 2013. Adaptive task assignment for crowdsourced classification. In *International Conference on Machine Learning*. 534–542.

[13] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical Report.

[14] Shweta Jain, Ganesh Ghalme, Satyanath Bhat, Sujit Gujar, and Y. Narahari. 2016. A Deterministic MAB Mechanism for Crowdsourcing with Logarithmic Regret and Immediate Payments. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 86–94. http://dl.acm.org/citation.cfm?id=2936924.2936941

[15] Shweta Jain, Sujit Gujar, Satyanath Bhat, Onno Zoeter, and Y Narahari. 2018. A quality assuring, cost optimal multi-armed bandit mechanism for expertsourcing. *Artificial Intelligence* 254 (2018), 44–63.

[16] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 467–474.

[17] Sergey Karayev, Mario Fritz, and Trevor Darrell. 2014. Anytime recognition of objects and scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 572–579.

[18] David R Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*. 1953–1961.

[19] Eric P Kukula, Stephen J Elliott, and Vincent G Duffy. 2007. The effects of human interaction on biometric system performance. In *International Conference on Digital Human Modeling*. Springer, 904–914.

[20] Hongwei Li, Bin Yu, and Dengyong Zhou. 2013. Error rate bounds in crowdsourcing models. *arXiv preprint arXiv:1307.2674* (2013).

[21] Padala Manisha and Sujit Gujar. 2019. Thompson Sampling Based Multi-Armed-Bandit Mechanism Using Neural Networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2111–2113. http://dl.acm.org/citation.cfm?id=3306127.3332027

[22] Michel Owayjan, Amer Dergham, Gerges Haber, Nidal Fakih, Ahmad Hamoush, and Elie Abdo. 2015. Face recognition security system. (2015), 343–348.

[23] Shourya Roy, Chithralekha Balamurugan, and Sujit Gujar. 2013. Sustainable Employment in India by Crowdsourcing Enterprise Tasks. In *Proceedings of the 3rd ACM Symposium on Computing for Development (ACM DEV '13)*. ACM, New York, NY, USA, Article 16, 2 pages. https://doi.org/10.1145/2442882.2442904

[24] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. 2015. Best of both worlds: human-machine collaboration for object annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2121–2131.

[25] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in neural information processing systems*. 3856–3866.

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. (2015), 815–823.

[27] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 614–622.

[28] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2015. Deeply learned face representations are sparse, selective, and robust. (2015), 2892–2900.

[29] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.

[30] Long Tran-Thanh, Sebastian Stein, Alex Rogers, and Nicholas R Jennings. 2014. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence* 214 (2014), 89–111.

[31] Long Tran-Thanh, Matteo Venanzi, Alex Rogers, and Nicholas R Jennings. 2013. Efficient budget allocation with accuracy guarantees for crowdsourcing classification tasks. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 901–908.

[32] Sudheendra Vijayanarasimhan and Kristen Grauman. 2009. Multi-level active prediction of useful image annotations for recognition. In *Advances in Neural Information Processing Systems*. 1705–1712.

[33] Catherine Wah, Steve Branson, Pietro Perona, and Serge Belongie. 2011. Multiclass recognition and part localization with humans in the loop. In *2011 International Conference on Computer Vision*. IEEE, 2524–2531.

[34] Catherine Wah, Grant Van Horn, Steve Branson, Subhransu Maji, Pietro Perona, and Serge Belongie. 2014. Similarity comparisons for interactive fine-grained categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 859–866.

[35] Jie Yang, Thomas Drake, Andreas Damianou, and Yoelle Maarek. 2018. *Leveraging crowdsourcing data for deep active learning an application: Learning intents in Alexa*. 23–32.

[36] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014).

[37] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2014. Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543* (2014).