**KAUST Impact Acceleration Fund**

**Project Proposal Title: A Quality Filtering System for Massive Visual Data Collection Applications**

**PI name: Yehia Massoud**

**Background**

Give a brief background to the project, what is it you aim to achieve, and why it is significant, how it is currently being done if other groups are addressing the same question, and what are the advantages of your approach over this.

Equipped with multi-sensing intrinsic capabilities (e.g., high definition cameras, geo-location and orientation sensors, computational resources), smartphones did not only revolutionize the way how people communicate with each other, but also the way how they interact with their surrounding environment. With the exponential spread of these devices, online photo sharing services such as Flickr and Instagram are gaining more and more popularity. In fact, according to Forbes magazine, people worldwide upload an average of 1.8 billion digital images to the internet every single day.

This clear willingness of people to share photos with their peers along with the abundant deployment of static and mobile video recording devices (e.g., CCTVs, camera-equipped UAVs, and on-dash mounted cameras) have promoted many applications requesting image/video collection from the crowd of people and IoT devices without any pre-deployed dedicated infrastructure. Such examples include photo tourism, media reporting of real-time events, weather monitoring, disaster recovery, and public safety and emergency management []. These mobile crowdsourcing systems enable distributed complex image-based tasks to be executed by mobile IoT devices through open calls \cite{9474925}. Such devices can be asked by a given entity, e.g., local authorities requesting services from the crowd, to take pictures (manually or autonomously) about an ongoing event or to document a specific scene. For instance, in the case of fire, people near the impacted area can provide photos or videos to the fire department. Clear and reliable visual data need to be efficiently collected and transmitted as quickly as possible to enable comprehensible assessment and take adequate measures.

Such system models require multiple inputs from different heterogeneous sources including users with different mobile devices’ features. It is hard to guarantee the quality and reliability of the submitted images/videos. Moreover, in most of the cases, the stream of visual data contains redundant and erroneous information. In fact, it may happen that some of the stream sources contain duplicate images and even irrelevant ones. Indeed, one way to maximize the event coverage context could be by requesting more and more inputs for a more accurate assessment of the situation. This will probably enhance the context given that the uploaded data will cover many angles and provide a clearer idea about what is going on site. However, such concepts will also lead to extensive processing to select relevant images or videos in addition to an increase of the network communication bandwidth as well as the operation cost.

A problem with these crowd-based systems is how to select relevant images and abolish the redundancy while keeping only the most indicative submissions. There is a need to quantify the quality of the submitted data based on the accessible geographical and geometrical information including the device's orientation, position, and all related parameters of the built-in camera. From such data, one can infer where and how the photo/video is taken, and then only maintain the most useful and relevant ones. Such processes are required to speed up the decisions of the photo/video collector that needs to make after comprehending the received data. This proposal aims to design an innovative artificial intelligence based approach to classify and filter the received data stream submitted by the crowd and maintain high-quality non-redundant images/videos by jointly processing them and their corresponding meta-data.

There has been some content-based image processing work investigating this problem. For example, in \cite{8788735}, a photo coverage crowdsourcing mechanism to forward non-redundant photos in emergencies using delay-tolerant networks has been proposed. Their proposed schemes relies on establishing dynamic point of interests to filter duplicate photos. In \cite{7807311}, a generic participatory picture collection framework called CrowdPic is proposed. The method suggested choosing the most diversified pictures to obtain the maximum coverage based on P-tree hierarchical data structure. These content-based image processing techniques usually demand too much computational and communication resources at both user and server ends. On the other hand, existing solutions from description-based techniques either categorize photos based on user defined tags, or prioritize them by the GPS location~\cite{6005632}. Obviously, tagging each photo manually may discourage public participation and might be inaccurate. Moreover, GPS location by itself may not be sufficient to reveal the real point of interest. Even at the same location, smartphones facing different directions will have different views.

To ensure the reception of high quality visual data with low-bandwidth consumption, this project proposes to investigate a smart and automated technique to analyze the submitted data and discard false submissions to preserve only the ones that maximize the event coverage (i.e., the most diverse data associated to the same event)[[1]](#footnote-1). In this context, we consider the generic scenario where a visual content requestor, defined as the requestor, aims to collect images/videos associated to different services from the crowd. For example, a third-party server is running a service for different stakeholders collecting data for a weather monitoring application, traffic management, and damaged infrastructure monitoring. This project aims to develop a quality filtering system for image/video submissions based on a series of machine/deep learning models to treat the diversified submissions in a real-time manner. As a first step, we propose to represent the submissions using effective numerical representations that we identify as “embeddings”. This multi-modal processing embeds the submissions based on both, the combinations of images’ visual and semantic features, and their associated metadata (e.g., when and where the photo were taken). Hence, these embeddings (small size data with an order of magnitude of Kilobytes) are transferred to the server for processing to decide which submission (large size data with an order of magnitude of Megabytes) to be selected and finally uploaded to the servers to be utilized by the requester.

The quality filtering system consists of two phases: classification and data filtering based on the generated embeddings only. The classification aims to categorize the embeddings into the different classes according to the services managed by the requestor. Then, the data filtering, consisting in executing checking measures on the embeddings including reliability, redundancy, and resolution check, is performed to discard erroneous, duplicated, and low quality submissions. At the end, and once the verdict has been determined, the server requests the selected devices to transmit the full resolution images/videos corresponding to the selected embeddings.

**Impact plan and proposed activities**

Briefly describe the activities proposed, with a rough timeline for each activity with milestones and decision points. Show how the expertise and all the resources required to successful complete this explorative project are available and accessible.

Concisely, this project focuses on devising a generic framework encompassing a series of deep/machine learning models. Each one of them focusing on a specific task of the quality filtering system to be designed. The activities to be undertaken in this project can be summarized as follows:

**Task 1: Autoencoders for Embedding Generation**

* **Objective:** Develop a sophisticated autoencoder neural network taking as inputs the features of the visual data (photo or video) in addition to the corresponding metadata to generate numerical representations, defined as embeddings, encapsulating information of the submissions.
* **Procedure:** The autoencoder will be trained over different datasets associated to selected services handled by the requester, i.e., the server. The selected datasets will cover services such as damaged infrastructure monitoring, fire management, and traffic control.
* **Expected Outcome and Decision Point:** The output of this task is a neural network autoencoder generating numerical representations providing a high-level of accuracy in representing the visual data. In other words, the similarity between the features of two visual data, e.g., photos, must be accurately measured using the similarity of their corresponding embeddings. To this end, different metrics will be applied including cosine similarity and mean absolute errors.
* **Duration:** This task is estimated to be completed in the first four months of the project.

**Task 2: Quality Filtering System Design**

* + **Objective:** Develop the two phases of the quality filtering system: clustering and filtering based on the generated embeddings of Task 1. In this task, two machine learning models will be developed to process the numerical representations. The objective is to design an automated tool to classify the submissions into different categories and determine the set of submissions to maintain and the one to discard by the requestor, i.e., the server.
  + **Procedure:** The models will be trained on the same datasets used in Task 1 and an exhaustive method will be adopted to measure the efficiency of the unsupervised method for filtering the embeddings.
  + **Expected Outcome and Decision Point:** The output of this task is two cascaded machine learning models, the first one is supervised for classification, while the second model is unsupervised, offering a high-level of accuracy in classifying the embeddings and maintaining required submissions.
  + **Duration:** This task is expected to last four months.

**Task 3: APIs and Quality Filtering System Simulation**

* + **Objective:** Once the three AI-based models are successfully developed, Task 3 is dedicated to integrate the different models into a single framework.
  + **Procedure:** Task 3 drives towards the implementation of an API at the front-end of the system, i.e., at the level of the visual data collector (e.g., smartphone) where the trained autoencoder neural network, developed in Task 1, will be implemented. It also aims to develop a server-side API allowing the processing of the submitted embeddings using the machine learning models developed in Task 2 at the server level.
  + **Expected Outcome and Decision Point:** A simulator of the proposed visual data filtering system will be tested to validate the operation of the developed technology. The simulator must successfully operate and treat the considered types of visual data, i.e., photos and video frames and maintain a high-level of accuracy in filtering the submissions. Comparisons with a manually filtered dataset will be performed to measure the effectiveness of the develop system. The speed of the whole procedure will be also evaluated to estimate the prospective efficiency of the system when applied for real-time applications.
  + **Duration:** This task will consume the remaining four months of the project.

The timeline of the project and the different activities is given as follows:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Month**  **Task** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| **Task 1** |  |  |  |  |  |  |  |  |  |  |  |  |
| **Task 2** |  |  |  |  |  |  |  |  |  |  |  |  |
| **Task 3** |  |  |  |  |  |  |  |  |  |  |  |  |

This project will require the utilization of open image and video datasets that are available in the cloud. In the following, we provide the links of some datasets that will be used in this project:

* Dataset 1
* Dataset 2
* Dataset 3

Although they are covering different services, these datasets can be easily labeled and their data can be mixed together to generate a heterogeneous dataset to be used as an input for our framework. The ITL lab, led by PI Massoud, have access to 20TB of storage drive, namely the WahaDrive provided by KAUST. The storage capacity can be extend up to 100 TB if needed.

To run the AI models, computationally powerful GPU machines are required. The ITL lab, led by PI Massoud, has acquired strong computers having GPU …. . Moreover, the supercomputing Core Lab at KAUST offers extremely powerful GPU servers that can be used to train our AI-based modules if needed.

For this project, the whole system is software-based. Hence, there is no major resource issue to report. At this stage, no hardware equipment is required.

**Budget request and justification**

Please outline required budget with justification. Note the maximum time frame for translation projects applicable to this program is no more than 12 months. Complete Budget template.

**Intellectual Property**

Has any IP related to this proposal been disclosed to KAUST Innovation? Is the technology to be developed unencumbered by any prior intellectual property or agreements?

As per of our knowledge, this technology is not encumbered by any prior intellectual property. In this project, sophisticated AI modules will be developed as described in Task 1 and Task 2. The basic versions of the AI modules that might be utilized are open source. Therefore, we do not expect any Intellectual Property conflict.

No IP related to this proposal has been disclosed to KAUST Innovation.

**Deliverables and definition of a successful project**

Describe what you expect and plan to have delivered by the end of the project term, what will be your mid-term deliverables to show the project is on track. What would be the next step after a successful project? Who will use the outcomes of this project and how will they use them? What actions have been taken or will you take during the project to prepare and plan for the next step?

The project is designed to output a rich menu of deliverables with respect to its duration (12 months). The list includes the following:

* A conference paper of 4-5 pages covering the outcomes of Task 1 to be delivered at Month 4,
* A mid-term report summarizing the progress of the project will be delivered at Month 7,
* A journal paper of 8-10 pages covering the outcomes of Task 2 to be delivered at Month 10,
* An end-user and server side APIs to simulate the proposed quality filtering system to be delivered at Month 12,
* A final report summarizing the achievements and outputs of the project to be delivered at Month 12.

At the completion of this project, the software Technology Readiness Levels (TRLs) rating is believed to fall within a range of 4 - 5. At this stage, the core of the system will be functioning on specific services and ready to be upgraded to a prototype operational in a real-world environment. To this end and as a future extension of this project, we plan to convert the framework into a complete and qualified mobile crowdsourcing system for various crowd-engaging services by:

* Applying transfer learning techniques on new datasets to enlarge the application domains of the proposed framework and accommodate it to more services aligned with the Kingdom interests.
* Integrating online learning mechanisms to ensure the continuous updates of the datasets and AI models of the framework.
* Focusing on reducing the computational complexity of the framework in the post-training phase to enable fast processing of the submissions and allow its application to delay-intolerant applications.
* Developing user-friendly software tools, i.e., mobile application and back-end dashboard. The “app” aims to allow end-users to submit their photos and control their contributions while the back-end allows the requestor to visualize the process and statistics.

The developed platform is designed to be a third-party platform interconnecting many stakeholders specialized in different services and soliciting the power of the crowd to gather informative data with omnipresent IoT devices. It also provides an automated tool for these stakeholders to speed up the data processing before final reporting of the collected data. In addition to that, it aims to relax the bandwidth congestion of the communication networks by operating on numerical representations of the data instead of the visual content itself. It is expected to significantly reduce the amount of data transfer and processing for these types of services. The ultimate outcome of the next steps of this project is its commercialization if its technical efficiency and the market analysis are positively favorable.

**What are the risks? Any other risk; not necessarily** technical

Describe the risks to the project’s success, and what you will do to address and minimize these risks.

The first task is instrumental for the success of the project. Achieving satisfactory results in Task 1 (determining an AI-based technique to represent visual content data with numerical vectors) allows the project to culminate all its goals. If not, Task 2 and Task 3 are still active and can be applied on the visual content of the data instead of the numerical ones. Therefore, the PI will hire a Researcher essentially specialized in the field of Task 1 to maximize the efforts spent on this Task.

The project is investigating two types of data to be processed with the AI modules: images and videos. We do not expect issues with the images. However, although they can be seen as a sequence of images, videos might impose some technical issues while training the AI modules. As an example, the training module might require a lot of time to converge. Therefore, we will resort to use the most powerful machines at KAUST provided by the supercomputing CoreLabs to maximize our computational capabilities. Otherwise, the system will continue working using images and more systematic and deterministic approach will be adopted for the videos.

1. In this project, we identify the visual data, namely images or videos, by the term “submissions”. [↑](#footnote-ref-1)