Offloading to Neighbouring Nodes in Smart Camera Network

Saurav Sthapit*, James R. Hopgood*, Neil M. Robertson† and John Thompson*

*Institute of Digital Communications, School of Engineering, University of Edinburgh
Emails: {s.sthapit, james.hopgood, john.thompson}@ed.ac.uk
†Vision Lab, School of Engineering and Physical Sciences, Heriot-Watt University
Email: n.m.robertson@hw.ac.uk

Abstract—Mobile Cloud Computing refers to offloading computationally intensive algorithms from a mobile device to a cloud in order to save resources (time and energy) in the mobile device. But when the connection to the cloud is non-existent or limited, as in battle-space scenarios, exploiting neighbouring devices could be an alternative. In this paper we have developed a framework to offload computationally intensive algorithms to neighbours in order to minimise the algorithm completion time. We propose resource allocation algorithms to maximize the performance of these systems in real-time computer vision applications (drop less targets). Results show significant performance improvement at the cost of using some extra energy resource. Finally we define a new performance metric which also incorporates the energy consumed and is used to compare the offloading algorithms.

Index Terms—Offloading, Mobile Cloud Computing, Energy

I. INTRODUCTION

Off-the-shelf smartphones are becoming ubiquitous and powerful, making them an interesting prospect to form a smart networked camera. However, they are not powerful enough for many applications, especially if the results are required in real-time. We have previously considered their pros and cons for distributed person re-identification [1]. If smartphones are deployed to carry out computationally intensive computer vision tasks, such as person tracking and re-identification between multiple cameras, they may not always be able to process everything within a user specified time. As such we can define performance of a system as the ratio of number of jobs processed to the number of jobs available.

Conventionally, computationally intensive algorithms have been offloaded to the “cloud” and it has been shown in some cases to save time and energy [2] [3]. In this paper however, we present a novel framework to offload these tasks to neighbouring mobile nodes which can significantly increase the performance without substantially depleting battery resource compared to the non-offloading case. We also present a single metric called Efficiency Score (ES) which also incorporates the energy consumption along with the performance.

A. Computing platform types

If there is no network connectivity, the only option is to do on-board processing. However, if there is some connectivity, we have the option to offload. We use the term “onloader” for the system which the “offloader” offloads its workload to. Some on-board processing can reduce the amount of data to be communicated while freeing up the onloader’s resources. For example, a background subtraction algorithm can limit the sensor from sending images with little or no activity. This saves communication cost for the sensor and the onloader has fewer jobs to perform. However, when the algorithms are fairly complex, the cloud is the preferred option. High round-trip latency can be compensated by its shorter runtime owing to high performance computing resources in the cloud. More details on mobile cloud offloading can be found in [4].

Neighbouring nodes are important alternative to the cloud for two reasons, the first reason is communicating with neighbours can help in the co-ordination and control of the node network. Neighbours can provide a cue of an incoming target, or give complementary and valuable information about the targets. More details about co-ordination and control can be found in [5]. The second reason is that when the cloud is unavailable, they can help in sharing the computational load. For example, in an underground transportation network, battle-space scenario, or a search and rescue mission after natural disaster, the internet may be unreachable. Even if there is a connection, the Network Bandwidth (BW) may be too low or intermittent. Neighbouring devices lack the computational power and energy of the cloud but may be readily available with high BW connections. However, we need to consider their available energy resources before offloading a job as when in the field, charging may not be readily available. In this paper, we discuss its feasibility in terms of time and energy.

Magurawalage et al. have considered offloading to intermediate cloud like entity called cloudlets [6] but to the best of our knowledge, battery powered neighbours have not been used by anyone as an offloading candidate.

B. Problem formulation

Let us assume there are two similar smartphones (X and Y) deployed in the field to identify people arriving in their Field Of View (FOV). Each can only process $N$ targets at a time. When there are $P > N$ potential targets to process, as shown in the Fig. 1 where $N = 3$, X has no other option than to offload or drop some of the targets. If the cloud is available and the bandwidth is sufficiently high, the cloud can be the...
onloader. Then, X can offload \((P - N)\) targets to the cloud and as Kumar et. al explains, it can save time and energy [2]. If the cloud is not available, conventional system would simply drop the targets. However, device Y has no target in its FOV at this moment. We show that neighbour devices like Y can be good alternative and may be used for offloading.

We now describe the simulator we developed to study smartphones with cellular and Wi-Fi communication capabilities. Then in section III, we describe how we use the simulator for a person re-identification system. Then we describe experiments in section IV and show results. Finally in section V, we present conclusions to the paper.

II. Modelling Network of Smart Cameras

The simulator allows us to use a simplified model of the algorithm flow for the target platform and update components easily as required. Wu et al. used queuing model theory to simulate workload on distributed nodes [3]. However, their assumption that when there is no workload the nodes do not consume any energy is not valid in real life. The major elements of our simulator relate to the algorithmic tasks, the sensor architecture, communication links and the targets. We go through each one in detail below.

A. Algorithmic tasks

The simulator’s model for the algorithmic task is characterised by its number of Operations (OP), input and output data size. For example, a person detection algorithm takes an image of size \(M \times N\) as the input, requires approximately \(C\) OP per image and outputs the number of persons in the image. Assuming one OP per clock cycle, we can estimate the execution time on the device using the clock frequency.

\[
T_{\text{exec}} \propto \frac{C}{\text{Clock Frequency}} \quad (1)
\]

We are aware that in different processors some OP take more than one cycle and multiple OP can be possible in one cycle, however this approximation (Eqn. (1)) gives an estimate of time required without detailed execution information. If desired, algorithms could be executed on the Device Under Test (DUT) to measure the execution time more precisely.

The number of OP required for an algorithm can change. For instance, in the Mixture of Gradients (MOG) algorithm for background subtraction, it depends on how quickly a matching Gaussian distribution is detected for the particular pixel [7]. To make calculations easier for the simulation, we take the worst case scenario where the matching Gaussian is not found.

B. Component Based Sensors

In order to realistically emulate its behaviour, a sensor is divided into its components such as the Central Processing Unit (CPU) and cellular radio. We do not consider the energy consumption by the display as it can be turned off by the application. We use the utilisation based model by Jung et al. to calculate the energy consumption [8] and our parameters are based on a Google Nexus I phone which was one of their DUTs. However if desired, the simulator can be calibrated for a different DUT in a straightforward manner.

\[
P_{\text{exec}} = P_{\text{idle}} \times (T_{\text{frame}} - T_{\text{active}}) + P_{\text{active}} \times T_{\text{active}} \quad (2)
\]

1) Image Sensor: The image sensor consumes significant energy in a mobile device when used continuously. According to Likamwa et al., the energy consumption per frame of the image sensor can be modelled as follows [9].

\[
P_{\text{camera}} = P_{\text{idle}} \times (T_{\text{frame}} - T_{\text{active}}) + P_{\text{active}} \times T_{\text{active}} \quad (2)
\]

where \(T_{\text{active}} = \frac{\text{Number of Pixels}}{\text{Camera Clock Frequency}}\). Based on Eqn. (2), we can either reduce the image resolution, thereby reducing \(T_{\text{active}}\) or reduce the acquisition rate to save the energy consumption.

2) Application Processor (AP): The CPU power is made up of two parts, idle power and the running power, as follows:

\[
P_{\text{cpu}} = \beta_{\text{freq}} \times t + \beta_{\text{idle}}. \quad (3)
\]

where \(u\) is the utilisation and \(\beta_{\text{freq}}\) and \(\beta_{\text{idle}}\) are the CPU parameters listed in Table I. We calculate the utilisation as the ratio of the CPU time used vs the time available per frame. However, the CPU is also used by the Operating System (OS) and other running applications. Darge used normal and exponential distributions to simulate workload in [10]. We also used a random variable \((r)\) sampled from a Gaussian distribution to simulate these other activities. By adjusting the mean of \(r\) we can simulate a busy sensor and idle sensor. The total utilisation is calculated as shown below.

\[
u = \frac{\sum_{i=1}^{N} T_{\text{exec}}}{T_{\text{frame}}} + r \quad (4)
\]

where \(N\) is the number of algorithms to be processed, \(T_{\text{exec}}\) is the execution time for \(i^{\text{th}}\) algorithm (see Table II for execution times for all algorithms) and \(T_{\text{Frame}} = \frac{1}{F_{\text{Frame}}}\) is the time available for each frame. In the situation where \(T_{\text{exec}} > T_{\text{Frame}}\) which is very likely in the case of algorithms for person re-identification; we only run the CPU to 100% load and run the remainder of the algorithm in the next frame and so on.

3) Cellular (3G): Cellular radio is modelled as a three state system: IDLE, Forward Access Channel (FACH) and Dedicated Channel (DCH). The IDLE mode is the non communicating mode and has the lowest power consumption. In this mode, the User Equipment (UE) is turned on but has not established Radio Resource Control (RRC) connection with the Radio Network Controller (RNC). In DCH state the UE has a dedicated transport channel for data transmission in both directions, but this is 50 to 100% more expensive than FACH, where FACH is the intermediate state with reduced power consumption and low data rate. There is no dedicated channel allocated in this mode and it can only transmit user data through shared low speed channel that is typically less than 15kbps [11]. As we can see from Eqn. (5), power is only dependent on state but not on utilisation. Fig. 2 shows the state diagram with the inactivity timers which along with data buffer size controls the state promotions and demotions.

\[
p_{\text{3G}} = \begin{cases} 
  \beta_{\text{IDLE}} & \text{if RRC state is IDLE} \\
  \beta_{\text{FACH}} & \text{if RRC state is FACH} \\
  \beta_{\text{DCH}} & \text{if RRC state is DCH} 
\end{cases} \quad (5)
\]

where RRC is the current state of UE and \(\beta_{\text{IDLE}}, \beta_{\text{FACH}}\) and \(\beta_{\text{DCH}}\) are based on [8].
Background subtraction is the basis of this research. In the connected mode, we ignore it as we assume a connection between the DUT based on [8]. If the number of packets per second exceeds the threshold of 20 then Wi-Fi is in the high power state, else in the low power state. Unlike the cellular system, the power consumption is directly proportional to the data rate. Although Wi-Fi consumes energy in scanning mode, we ignore it as we assume a connection between the sensors as the basis of this research.

III. SYSTEM DESIGN

In this section, we show how offloading may be used to increase the performance of a system. We consider a pedestrian re-identification system outlined in Fig. 3. It starts with image acquisition from the image sensor. A background subtraction and a person detection algorithm is applied on the image to detect the number of people in the view. When a pedestrian is detected, we apply a person re-identification algorithm to each detection such as [12]. Our goal is to identify as many detections as possible.

A. Application Partitioning

The algorithmic complexity of the person re-identification algorithm outweighs that of other algorithms in the chain (see Table II). So, the overall complexity of the system can be estimated as \( O(N) \) where \( N \) is the number of people detected. To be realistic, we limit the number of people in an image \((800 \times 600)\) in the simulator to be fewer than 10.

B. Energy Saving Methods

We replicate following energy saving techniques to make the simulation realistic as much as possible.

1) Dynamic Frame per Second: We can save energy by decreasing the number of FPS of the system (see Eqn. (4)). However, very low FPS may mean some of the detections may be missed. We implemented an algorithm to vary the FPS of each individual sensor between 1 and 16 in the following way.

\[
\text{FPS (new)} = \begin{cases} 
\text{FPS (old)} \times 2 & \text{if } t < \tau \\
\text{FPS (old)} \div 2 & \text{if } t > \tau 
\end{cases} \tag{7}
\]

where \( t \) is the time between target activities and \( \tau \) is 5 seconds.

2) Dynamic Voltage and Frequency Scaling (DVFS): A simple algorithm controls the clock frequency of the sensor. When the CPU utilisation is below 0.4, the clock frequency is lowered according to Table I and it is scaled to maximum frequency as soon as the utilisation is above 0.9.

C. Offloading

We classify only the re-identification algorithm as offloadable as for others, the communication costs and the time delay outweighs the benefits of offloading. Offloading an algorithm entails sending input data, waiting for the onloader to execute, and receiving output data. Before transmitting however, the data has to be formatted in packets and some overhead will be added to the processor. These operations can be a few hundred per packet which needs to be added to the CPU workload.

1) Time Cost: The communication times are proportional to the data size to be communicated and inversely proportional to the network BW. We assumed the BW to be static. Waiting times can be estimated using Eqn. (1) for the onloader’s clock frequency. But, it does not take into account the CPU load. If our onloader’s AP is already busy, our estimation can be very far from reality. So we re-write the time calculation using onloader’s average CPU utilisation.

\[
T_{\text{wait}} = \frac{T_{\text{exec}}}{1 - \mathbb{E}[u]} \tag{8}
\]

where \( \mathbb{E}[u] \) is the average of \( u \) from Eqn. (4). The total time cost which is also known as makespan, is shown below [13].

\[
T_{\text{total}} = T_{\text{packet}} + T_{\text{send}} + T_{\text{wait}} + T_{\text{receive}} \tag{9}
\]

where \( T_{\text{packet}} \) is the time to format the data in a packet.

2) Energy Cost: There are two energy costs involved. The first is for the offloader (\( E_{\text{off}} \)) and includes data packeting and the radio communication cost.

\[
E_{\text{off}} = (T_{\text{send}} + T_{\text{receive}}) \times P_{\text{radio}} + P_{\text{cpu}} \times T_{\text{packet}} \tag{10}
\]

where radio is \( \{3G, WiFi\} \). Second is for the onloader which includes radio cost, execution cost and the packeting costs. So far, in literature, cost for the onloader is ignored as energy is not of major concern for the cloud. But while offloading to the neighbours, we need to consider it.

\[
E_{\text{on}} = (T_{\text{receive}} + T_{\text{send}}) \times P_{\text{radio}} + P_{\text{cpu}} \times (T_{\text{execute}} + T_{\text{packet}}) \tag{11}
\]

D. Multi-Objective Optimisation

The time and energy costs from Eqn. (9, 10 and 11) can be inferred as variables of a multi-objective optimisation problem.

\[
\text{Cost} = w_{\text{time}} \times T_{\text{total}} + w_{\text{off}} \times E_{\text{off}} + w_{\text{on}} \times E_{\text{on}} \tag{12}
\]

where \( w_{\text{time}}, w_{\text{off}} \) and \( w_{\text{on}} \) are the weights for each objective. The cost function involves adding time and energy variables (i.e. different units), which requires careful selection of the
weights. We avoid this situation by limiting one of the objectives to a threshold (ε) and optimising rest of the objectives [14]. Regarding the real-time nature of our problem, we limit the time and optimise the energy variables. We set ε = 25 seconds and leave out the nodes that do not satisfy this constraint (denoted by the light dots in Fig. 4). It is still a multi-objective problem but only with two variables of the same unit (Joules). We now study three methods to optimize offloading performance.

1) Minimize Energy Cost (MEC): In this method, we choose the node that satisfies the time constraint described above and incurs the minimum offloader and the onloader energy cost. The solution is pareto-optimal and denoted by nodes on the line in Fig. 4 [14].

\[
\text{Cost}_{\text{MEC}} = w_{\text{off}} \times E_{\text{off}} + w_{\text{on}} \times E_{\text{on}}
\]  
(13)

2) Minimize Battery Impact (MBI): In a battery-powered device, using the least energy cost per job alone may not increase device lifetime. For example, say an algorithm requires 10 and 8 Joules on devices X and Y respectively. But X and Y have 500 and 50 Joules left in their battery respectively. Considering energy cost alone, Y is the best choice but when we consider the amount of energy left in the device clearly X is a better choice. We re-write Eqn. (13) as follows.

\[
\text{Cost}_{\text{MBI}} = w_{\text{off}} \times \frac{E_{\text{off}}}{E_{\text{rem. off}}} + w_{\text{on}} \times \frac{E_{\text{on}}}{E_{\text{rem. on}}}
\]  
(14)

where \(E_{\text{rem. off}}\) and \(E_{\text{rem. on}}\) are the energy left in the offloader and the onloader nodes.

3) Offload Only if Busy (OOB): The previous methods try to find the global solution, but offloading has overhead costs. So, this method tries to offload only if on-board processing is estimated to be infeasible. To do so, we add all the operations in the execution queue and use Eqn. (1) to estimate the minimum remaining processing time. If this time is greater than the threshold (ε), offload the algorithm minimising the time and energy objectives defined in Eqn. (12).

IV. SIMULATION AND RESULTS

We simulated a number of sensors connected to each other by Wi-Fi and to the server (when available) by cellular link. For simplicity, we assume that resource information about all the nodes (remaining energy, current CPU load etc.) is available and all the sensors have same computational capability but the server is 10 times more powerful. Also there is no energy limitation for the cloud so the weight \(w_{\text{on}}\) for the cloud is set to zero. Wi-Fi is set to 10 Mbps (high BW) whereas cellular is slower and set to 1 Mbps. At the start of the simulation, the battery level is uniformly distributed between 0 – 10 Watt-hour. The mean of \(r\) in Eqn. (3) is uniformly distributed from 0 – 1 (full load) and the standard deviation is fixed to 0.1. These parameters do not change during the simulation. Full simulation data and parameters are available here 1. Fig. 5 shows a snapshot of the simulation, where the blue squares and red dots represent the sensors and targets respectively. The targets are generated using a Poisson distribution and follow the random waypoint model [15]. In this model, targets move from one point to another with random speed and can also pause for random amount of time. When they enter into the FOV of the sensors (shown as green areas in Fig. 5), they are tagged as detected targets. The detected targets are then identified. Once the target has been detected and identified, it does not have to be re-identified again in the same camera. This is done to simulate tracking the target in each camera and using the best representation for identification purposes.

We tested the algorithms with various parameters using 100 Monte-Carlo runs each representing a 10 minute period. The results are listed in Table III. When the cloud is not available, Successful Identifications (SI) (which is targets detected minus targets dropped) improved from 10.4 in the Non Offloading (NO) case to 12.7 for the MEC case and 13 for OOB case but degraded to 8.87 for MBI. MEC and OOB boosted the performance by more than 20% while only incurring around extra 10% energy consumption. MBI did not perform well because of the communication overhead. However, if the runs were longer, may be we would see some improvement in the device lifetime. All the algorithms did better than the NO case when the cloud was available, dropping almost no targets but the energy consumption was significantly higher. This shows that offloading to cloud blindly may increase the performance short term but may shorten device lifetime.

\[^{1}\text{http://sauravsthapit.com.np/EUSIPCO2016/}\]
It improved the performance by approx. 25% while costing about 10 – 20% more energy. The ES metrics suggests that energy is used more productively. Contrary to the general belief, the results show that given the constraints in bandwidth offloading to the cloud may not be the best option in terms of performance and energy cost. In future work, we plan to evaluate more dynamic scenarios.

VI. Acknowledgement

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [EP/K014277/1] and the MOD University Defence Research Centre (UDRC) on Signal Processing. Simulation information available at 2

REFERENCES


2http://sauravsthapit.com.np/EUSIPCO2016/