VIDEO QUESTION ANSWERING FOR SURVEILLANCE

Muhammad Iqbal Hasan Chowdhury, Kien Nguyen, Clinton Fookes, Sridha Sridharan

Queensland University of Technology
{m2.chowdhury,k.nguyenthanh, c.fookes, s.sridharan}@qut.edu.au

ABSTRACT

There are many task in surveillance monitoring such as object detection, person identification, activity and action recognition etc. Integrating variety of surveillance task through a multimodal interactive system will benefit real-life deployment, and will also support human operators. We first introduce a dataset which is first of its kind and named as Surveillance Video Question Answering (SVideoQA) dataset. The multi-camera surveillance monitoring aspect is considered through the multimodal context of Video Question Answering (VideoQA) in the SVideoQA dataset. This paper proposes a deep learning model where VideoQA task on the SVideoQA dataset is attempted to solved in a manner where memory-driven relationship among appearance and motion aspect of the video features are captured. At each level of the relational reasoning respective attentive parts of the context of the motion and appearance features are identified forwarded through frame level and clip level relational reasoning module. Also, respective memories are updated which are again forwarded to the memory-relation module to finally predict the answer word. The proposed memory-driven multilevel relational reasoning is made compatible with the surveillance monitoring task through the incorporation of multi-camera relation module, which is able to capture and reason over the relationships among the video feeds across multiple cameras. Experimental outcome exhibits that the proposed memory-driven multilevel relational reasoning perform significantly better on the open-ended VideoQA task compared to other state-of-the art systems. The proposed method achieves an accuracy of 57% and 57.6% respectively for the single-camera and multi-camera task of the SVideoQA dataset.

***Index Terms*—** Visual Question Answering (VQA), Surveillance Monitoring, Relational Reasoning, Scene Understanding.

1. INTRODUCTION

Video Question Answering (VideoQA) is the task where natural language question is imposed on the arbitrary part of any video sequence, and relevant answer is automatically generated by the machine. The image counterpart of this task is called visual question answering (VQA). This task is challenging as it involves addressing effectively the varying and unpredictable nature of natural language question in addition to the temporal dynamics of video contents. It is expected that solving this kind of problem will lead the machine to become more human-like in terms of logical and commonsense reasoning process.

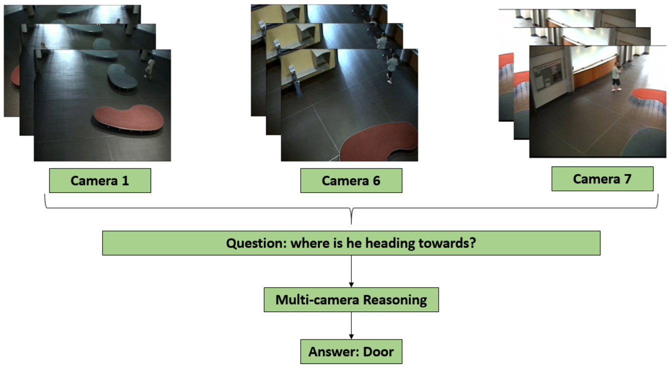


Fig. 1. VideoQA task on the surveillance video feeds require multiple camera inputs to get the overall question specific context. A single input video from a camera is not always sufficient to infer the correct context specific answer.

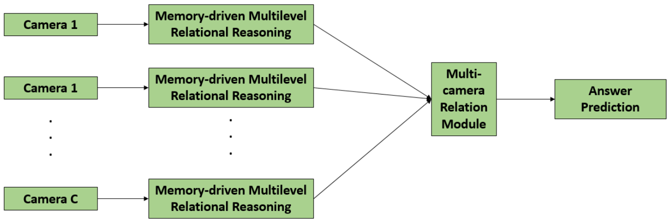

Camera 1
Camera 1
Camera C
Memory-driven Multilevel Relational Reasoning
Memory-driven Multilevel Relational Reasoning
Memory-driven Multilevel Relational Reasoning
Multi-camera Relation Module
Answer Prediction

Fig. 2. Each of the camera feeds are passed through the memory-driven multilevel reasoning as shown in Figure 3, when a single natural language question is imposed on several video feeds. Outputs of each of the memory-driven multilevel reasoning is considered as input to the multi-camera relation module, where question conditioned relation among multiple cameras are derived.

VQA [1] and/or VideoQA [2] combines both the text and visual (image/video) modality in the common context of reasoning. A natural language question is imposed on the respective image/video to generate the relevant answer. However, this task not only generate/predict an answer rather it forces the machine to perform human-like understanding and reasoning of data from two different modality. The VQA task also shows the progress of the machines' capability of multimodal understanding. Machine's capability of the multimodal understanding [?] is transferable to the the surveillance monitoring task. The use of VideoQA is surveillance monitoring [?] will make the monitoring system more interactive. Instead of manually searching over huge amount of videos, the respective user can simply put a question against the captured video data. As an example, let us consider a question like, "how many red cars were visible in the car park between 4am to 9pm exiting through the north gate?". The machine needs to perform object detection, counting, tracking etc. to generate all possible outcomes for such questions. Research of VideoQA system are thriving towards inventing models which can perform variety of the tasks poses by the natural language question in an end-to-end manner.

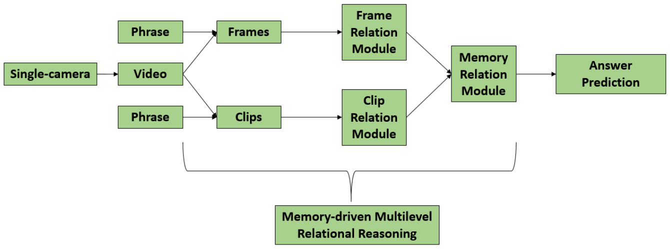


Fig. 3. This is the overall schematic of the reasoning process of the proposed method. Both appearance and motion based features are considered through frame-relation module and the clip-relation module. The subsequent memory-relation module finds and preserves the memory upate operations which is performed through the frame and clip-relation modules. The schematic operation of this diagram is performed in a phrase-by-phrase manner based on the posed natural language question.

The general VQA or VideoQA dataset are not directly suitable to be applied for the surveillance settings. One of the main reason behind this is the fact that there is no multi-camera annotations available in these dataset. Surveillance monitoring usually comprises of multiple cameras, and thus the annotation of dataset should also include video feeds and respective question-answer annotations to train models on the surveillance video question answering task. Also, surveillance monitoring requires to reason over both general activities and any anomalous activities. Thus, an ideal surveillance video question answering dataset should include instances of both trivial and anomalous activities with respective open-ended question-answer annotation. This paper proposes the SVideoQA dataset with multi-camera open-ended question-answer annotations, which is believed to be providing an interactive multimodal interface for automated surveillance monitoring. Figure 1 shows the importance of the inclusion of the multiple cameras from the surveillance. Output of three different cameras are required to effectively answer the question: "where is he heading towards?". The question is asked about the target person involved in the captured footage. The video feeds from any single camera is not sufficient to understand the overall intention or movement of the target individuals. So, video feeds from multiple cameras are required to get the overall picture of the target person person's intention, which is required to answer the posed question. Details about the proposed SVideoQA dataset is described in section 3

Existing VideoQA approaches in the literature do not deal with multi-camera reasoning, hence we have to propose a new one. Traditional single-camera VideoQA approaches rely on performing attention operation on video and textual features. Most of the approaches consider the word tokens of the question sentence one-by-one and calculate gradient based soft attention over the video features. In [3], each to the question sentence tokens are considered one after another to derive the attention and the video representation based on which the final answer word is predicted. Both appearance and motion based features are considered in the model of [3]. Again, in [?], a dynamic memory network [4] based approach is used where an episodic memory [?] is updated based on appearance and motion facts, which are obtained respectively from appearance and motion based video features in multiple cycles. Existing approaches uses attention mechanism and memory networks to tackle the challenges of video question answering on a single-camera settings. The application and usefulness of the video question answering can be further extended with the inclusion of multi-camera video feeds. Multi-camera video feeds are quite obvious in case of surveillance monitoring where the video question-answering can provide more automated instinct of what is happening across video feeds in a multi-camera network.

In this paper, we propose a multilevel reasoning process based on the question sentence. The proposed methods includes both single-camera and multi-camera video feeds to perform the reasoning conditioned on the phrases of the imposed question sentence. Models are trained for the single-camera and multi-camera video feeds in an end-to-end manner. Figure 3 shows the overall architecture of the proposed method for a single-camera setting. The video feed for a single-camera are divided into frames and clips, which are then forwarded respectively to the frame-relation module (FRM) and the clip-relation module (CRM). The outputs of the FRM and CRM are then forwarded through the memory-relation module (MRM) where relationships among updated memory vectors are found. An additional multi-camera relation module (MCRM) is included in case of the multi-camera task, where FRM, CRM, MRM and MCRM are jointly trained in an end-to-end manner to predict the right answer based on the open-ended question and video feeds from multiple-cameras. The multi-camera scenario of the proposed method is shown in Figure 2. The memory-driven multilevel operation is the same as shown in Figure 3. The MCRM is used to stitch the reasoning over multiple camera feeds in the single context of an imposed open-ended natural language question. The proposed method uses the principle of relation-network [5] which is also effective in performing visual commonsense reasoning [?] on videos.

Experimental results demonstrates the competitiveness of the proposed model with the state-of-the art approaches for single-camera VideoQA dataset. The proposed approach is able to capture the context of the question sentence effectively in a multi-level manner with the help of the memory-driven relational reasoning. With this architecture we are able to capture the question sentence context in a manner, where the evolving context of the question is better captured either on a single or multi-camera video network.

2. RELATED WORKS

2.1. Surveillance Monitoring

There are many efforts to automate several aspects of video surveillance. There are improvement in crowd counting [?], re-identification [?], camera calibration [?], abnormality detection [?] [?], event recognition [?], object tracking [?] etc. These efforts results in many commercial video analytic solutions. There are many system now which can robustly track an individual once instantiated by a human operator. These highly capable video management systems are now equipped with low-level image processing tools e.g. perimeter intrusion detection, loitering and abandoned object detection. But the fact is still these system are not truly interactive with human users which could enhance effective and robust monitoring in real time. There are attempts to retrieve events [?] which requires an annotated model to be learnt or searching people with specific criteria. But the fact is these endeavours not actually answers very natural intuitive query for the human investigator.

Surveillance camera monitoring has been increasing vastly which results in huge amount of visual data generated in each minute. Traditional style is to evaluate these data in a forensic mode after something has happened. Real time interactive monitoring and data analytics is not yet achieved. This manual analysis is labor intensive and error prone. Drawbacks of present system are that these are lack of naturally interactive, automated and scalable monitoring.

Several challenges are associated with an intelligent surveillance system with heterogeneous information e.g. quality of CCTV data, uncertainty of recognized events, inconsistency or conflict among multiple sources, adequate modelling of events information, composition of elemental events, scalability of the system, building ontologies for surveillance system etc.

A large scale video dataset designed for real-world surveillance event detection is proposed in [?]. A discussed in [?] traditional datasets for action recognition are not appropriate for real-world surveillance since those are consists of short clips showing each and every action by one individual [?] [?]. Again other dataset e.g. movies [?] and sports [?] do not depicts the aspect of surveillance in general. VIRAT Video Dataset is introduced in [?] is an expectation to boost further research in continuous visual event recognition (CVER).

2.2. Visual Question Answering

Visual question answering (VQA) refers to both image based and video based question answering. Video question answering (VideoQA) evolves as a natural extension of the image question answering (ImageQA) task. Both image and video based question answering have different forms including, 'open-ended', 'multiple-choice' and 'fill-in-the-blanks'. This paper uses datasets where 'open-ended' questions are being imposed on video feeds either from single or multiple cameras.

The straight forward approach to solve the VQA task is to extract image features through convolutional neural network (CNN), and the question sentence is encoded with a recurrent neural network. These are similar to approach which is proposed in [6]. Attention mechanisms [7] are proposed to allow VQA model to focus on specific regions of visual features. Generally attention mechanism refers to focusing on specific image regions. However, success in attention mechanism for ImageQA leads to use it in the VideoQA task too, which involves both static and temporal aspect of the feature representation. Stacked attention network [8] shows significant performance improvement for the ImageQA task. Also, a question-word guided attention [9] is proposed to solve the ImageQA task more effectively. Again, attention mechanism applied with compact bilinear pooling [?] shows significant improvement in accuracy to solve the ImageQA task. Successes in attention mechanism applied in ImageQA task evolves in using similar attention-based techniques to sovle the VideoQA task. However, attention mechanisms for videoQA need to address the temporal aspect of the video features representation.

Videos consist of sequence of static events, which are temporally coherent in nature. Video provides diverse and varying context which are usually sparsely distributed. An open-ended may refers to any part of the video. Thus, information from a single static image is never enough to correctly predict the answer word. VideoQA models need to narrow down its reasoning scope based on the question sentence. In addition with defining the context of the question, VidoeQA models need to perform attention based reasoning in multiple steps. A more realistic scene is to involve video features from multiple cameras, which is suitable for surveillance automation. The need to performing reasoning through multiple iterations leads to the use of memory networks [?]. An improved dynamic memory network [4] is successfully used to solve the ImageQA task. Use of both the attention mechanism and memory networks shows a new direction to address the VideoQA problem.

Previous ImageQA problems are extended in [2] and [10] to solve the VideoQA task. Generally successes and advancements in video/image captioning and attention mechanisms provide new research direction to solve the VideoQA task. An encoder-decoder based approach is proposed in [11], where unification of attentions is performed by considering both the quesiton sentence and the video. Frame-based visual attributes and question sentence based textual attributes are jointly learned in the approach proposed in [12]. An attention mechanism is proposed in[3], where attention based video representation is obtained from both appearance and motion based video features by sequentially considering each word of the question sentence. A motion-appearance co-memory network is proposed by [?], which is built on the concept of dynamic memory network [4]. In [?], author used the principle of relation network [5] mechanism to better capture the visual common sense knowledge in videos. However, [?] does not use the any memory equipped method to address the question answering problem. Also, the principle of relation network is used in [?], but their approach does not include any memory update operation. Also, the question sentence is processed in a word-level manner which results in the loss of the context of the question sentence during the reasoning process. A heterogenous memory update opertion is proposed in [?] for the videoQA problem. The proposed method in this paper adopts the memory-update operation described in [?] with the multilevel relation modules to better perform the question specific reasoning operation.

However, previous approaches on VidoeQA largely focus on problem of single-camera VideoQA. Also, the effectiveness on considering the evolving context of the question sentence in a phrase-by-phrase manner is not explored for the VideoQA problem. The proposed method in this paper, successfully uses the memory-driven relation network mechanism to capture the evolving context of the question sentence in both single-camera and multi-camera setting.

3. SURVEILLANCE VIDEO QUESTION ANSWERING DATASET

The SVideoQA dataset annotation is performed on an existing video recording of a person re-identification dataset [?]. This person re-identification dataset has video recording of 150 target person moving in a building environment across eight different cameras. Videos are captured at 25 frame per second with a resolution of 704×576. The question-answer annotation is conducted on the recording of each camera of the person re-identification dataset. Table 2 provides example of the single-camera and multi-camera question-answer annotation from the proposed SVideoQA dataset. There are $5,049$ question-answer annotations are available for the single-camera setting of the SVideoQA dataset. Also, the multi-camera setting of the SVideoQA dataset has 881 question-answer annotations available.

Cameras with overlapping viewpoints	Region Name
C1, C6, C7	R167
C2, C5, C8	R258
C3, C4	R34

Table 1. Overlapping camera regions based on which the question-answer annotation is performed for the multicamera VideoQA task of the SVideoQA dataset.

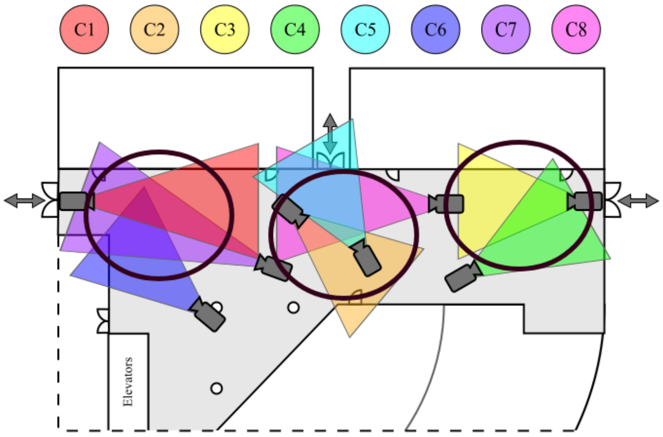


Fig. 4. Camera placement of the SVideoQA dataset, which identifies the overlapping regions covered by the multicamera viewpoints. Multicamera question-answers are annotated based on these overlapping regions. [?]. Circled region in the image shows the overlapping region of cameras. From left to right, the first, second and the third circles respectively refers to the regions R167, R258 and R34.

There are 152 video sequences available to perform the question-answer annotation. Each of the video sequence has eight distinct camera viewpoints. The question types of the proposed SVideoQA dataset fall into three borad classes, namely 'action search', 'group activity search', 'attribute search'. These three classes of question types covers the basic definition surveillance monitoring where behaviour, activities and information is captured and analyzed in the context of the available video footage. As shown in table 4, the 'walking direction'question types falls under the question class of 'action search'. The question types of 'counting'and 'walking companion'falls under the question class of ‘group activity search’. Again, the question types ‘person attribute’and ‘object use’fall under the question class of ‘attribute search’. Distribution of each of the individual question types are shown in Figure 5. In case of the multi-camera settings, there are higher number of ‘walking direction’types questions because walking across the viewpoint is the dominant activity in the current video recording which requires the video feeds from multiple cameras. Around 66.3% of the total question-answer annotations of multi-camera settings fall under the question type of ‘walking direction’, where *’counting’*, *’person attribute’*, *’object use’* and *’walking companion’* respectively cover 4.2%, 8.4%, 18.2% and 2.95% of the multi-camera question-answer annotations. Again, in the single-camera setting, there are higher percentage of ‘counting’type questions available. The ‘counting’type questions fall under the question type class of ‘group activity search’, and in the captured footage in many cases people are walking together, which contribute to a higher number of questions related to the ‘counting’type. 44.6% of the total single-camera question-answer annotation fall under the ‘counting’question type. Also, *’walking direction’*, *’person attribute’*, *’object use’* and *’walking companion’* respectively cover 10.95%, 26.38%, 15% and 3.01% of the total single-camera question-answer annotation. The multi-camera setting of the SVideoQA dataset focuses on the movement of the subject across camera viewpoints. Camera placement of the person re-identification dataset [**?**] is considered to annotate the question-answer for the multi-camera VideoQA task. Figure 4 shows the mapping and placement of the camera, which is used to capture the dataset [**?**] for the person re-identification task. It is visible from the Figure 4 that there are multiple overlapping regions, which holds the context of specific moments of the movements of the respective subjects. The circled areas in Figure 4 shows the overlapping regions of camera viewpoints. Table 1 holds the region name of the overlapping camera viewpoints, which is similar to be found in the proposed SVideoQA dataset annotation. Multi-camera question-answer annotations of the SVideoQA dataset is performed based on these overlapping regions.

Annotated questions are categorized into four different groups namely, *’walking direction’*, *’counting’*, *’object use’* and *’walking companion’*. Figure 5 provides percentage of each of the question types according to the annotation of the SVideoQA dataset. *’Walking direction’* type includes questions about on which side the target person is intended to take a turn. Again, *’counting’* type of questions ask about the number of visible people or stationary facilities like door, benches etc. Also, *’object use’* type questions ask questions about the objects (bags, sunglass, cellphone etc.) which are being carried or used by the target person. In addition, *’walking companion’* type questions ask about the person who is accompanying the target person, and this type of question mainly ask about the gender of the accompanying person. The proposed SVideoQA dataset is available for download[1].

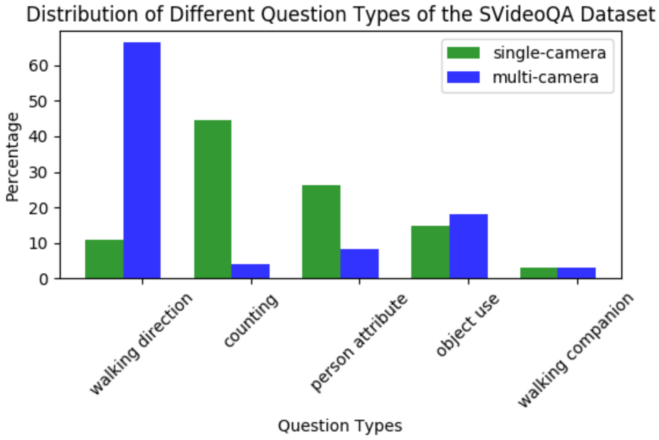


Fig. 5. Distribution of different question types of the SVideoQA dataset. The green and blue bar respectively refers to the single-camera and multi-camera settings of the SVideoQA dataset.

4. MEMORY-DRIVEN MULTILEVEL RELATIONAL REASONING

In this chapter, we propose a multilevel reasoning process based on the question sentence. The proposed methods include both single-camera and multi-camera video feeds to perform the reasoning conditioned on the phrases of the imposed question sentence. Models are trained for the single-camera and multi-camera video feeds in an end-to-end manner. Figure 3 shows the overall architecture of the proposed method for a single-camera setting. The video feed for a single-camera is divided into frames and clips, which are then forwarded respectively to the frm and the crm. The outputs of the FRM and CRM are then forwarded through the mrm where relationships among updated memory vectors are found. An additional mcrm is included in the case of the multi-camera task, where FRM, CRM, MRM, and MCRM are jointly trained in an end-to-end manner to predict the right answer based on the open-ended question and video feeds from multiple cameras. The multi-camera scenario of the proposed method is shown in Figure 2. The memory-driven multilevel operation is the same as shown in Figure 3. The MCRM is used to stitch the reasoning over multiple cameras feeds in the single context of an imposed open-ended natural language question. The proposed method uses the principle of relation-network [5], which is also effective in performing visual commonsense reasoning [**?**] on videos.

$$Answer, \mathbf{A} = Model(\mathbf{Q}, \mathbf{F}) \tag{1}$$

---

[1] SVideoQA Dataset download link: https://bit.ly/2L1Y9GX

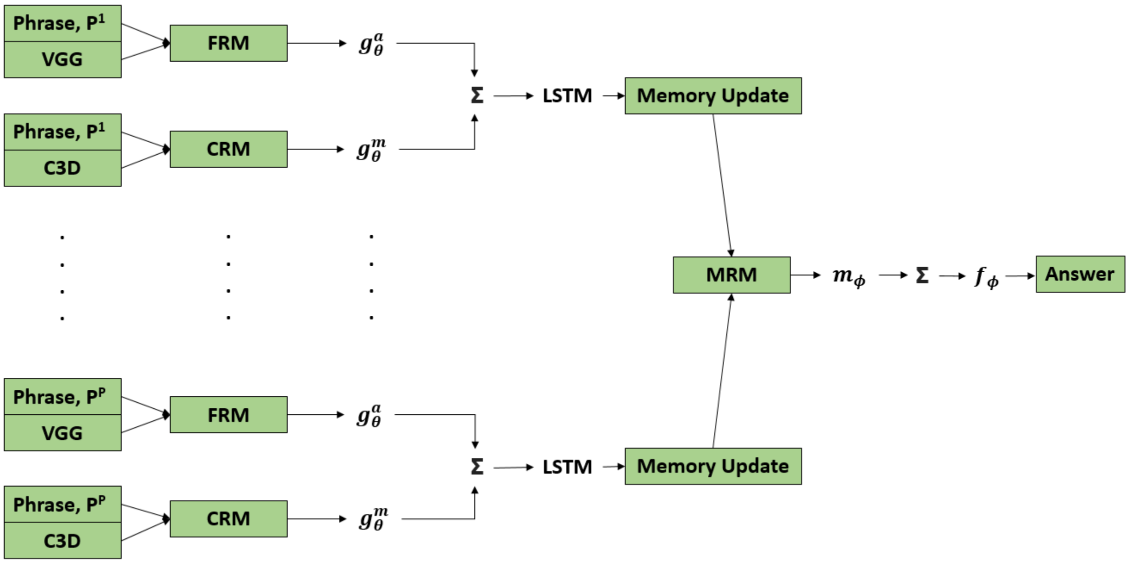


Fig. 6. If a question Q consists of P number of phrases then each of the phrases are considered to associate motion and appearance features to be passed respectively through the FRM and CRM. Outputs of FRM and CRM are passed thorough multilayer perceptron g_θ^a and g_θ^m. Output of multlayer perceptrons are passed through and LSTM before making the respective memory update for each of the phrase associated outputs of FRM and CRM. The P number of memory updated outputs are passed an input to the MRM, which results in finding the relationship among memory updates. Each of the outputs of MRM are then forwarded to the multilayer perceptron m_ϕ, which are summed together before forwarding to the multilayer perceptron f_ϕ to predict the final answer word.

A multilevel reasoning process is proposed in this chapter, which is done according to the context of the question sentence. The proposed reasoning is conditioned on the phrases of the question sentence for both single-camera and multi-camera video feeds. The proposed model is trained in an end-to-end manner for both the single-camera and multi-camera annotation of the SVideoQA task. Figure 3 shows the overall architecture of the proposed method for a single-camera setting. The video feed for a single-camera is divided into frames and clips, which are then forwarded respectively to the frm and the clip-relation module (CRM). The outputs of the FRM and CRM are then forwarded through the memory-relation module (MRM) where relationships among updated memory vectors are found. An additional multi-camera relation module (MCRM) is included in the case of the multi-camera task, where FRM, CRM, MRM, and MCRM are jointly trained in an end-to-end manner to predict the right answer based on the open-ended question and video feeds from multiple cameras. The multi-camera scenario of the proposed method is shown in Figure 2. The memory-driven multilevel operation is the same as shown in Figure 3. The MCRM is used to stitch the reasoning over multiple cameras feeds in the single context of an imposed open-ended natural language question. The proposed method uses the principle of relation-network [5], which is also effective in performing visual commonsense reasoning [**?**] on videos.

The key point of reasoning for the VideoQA task starts by analyzing the context of imposed natural language questions. A single-camera task requires to reason only a single sequence of frames, but for the multi-camera task, the frame sequences from multiple cameras need to be taken into consideration. The video, i.e. the frame sequences, are a continuous flow of visual context, which carries a lot of information. The question sentence defines the context, based on which, the VideoQA model needs to perform the required reasoning to derive the answer word. This procedure can be formulated in general as shown in Equation 1, where Q is the imposed question and F is the frame sequences i.e. the video. $Q = [x_{tok1}, x_{tok2}, ..., x_{tokN}]$ and $F = [f_1, f_2, ..., f_N]$ are both input to the video QA model which performs the necessary reasoning process to derive the answer word A.

The temporal features in the form of motion are an integral part of the reasoning in the VideoQA task. Also, the static aspect, i.e. the appearance-based features, play a crucial role in the reasoning process. Based on the question, either or both of the appearance and motion-based features are needed to be taken into consideration. Initially, the posed question sentence is divided into phrases, which is the meaningful chunk to drive the subsequent reasoning operation for both the single-camera and multi-camera VideoQA tasks. Let us consider question Q, which consists of P number of phrases. The proposed model will start by considering the $phrase1$

first, and gradually all subsequent phrases will be taken as input with the appearance and motion-based features.

FRM and CRM modules are responsible respectively for handling the appearance and motion-based features along with question sentence phrases. The outputs from the FRM and CRM modules drive the subsequent memory update operation. Finally, the output of the memory update operation for each of the phrases is fed into the MRM, which leads to the prediction of the final answer word. The proposed method combines multiple levels of relation modules with the memory update operation, and the operation is conditioned on the respective phrase of the question sentence. Firstly, the phrase with appearance and motion features are considered through FRM and CRM, which is the first level of the relation module. The output of FRM and CRM are passed through long short-term memory (LSTM), which is then fed to the memory update operation and the updated state of the memory is passed through, combined with the memory updated states of the phrases through the MRM module. The MRM is another level of relation module. The scenario described in this paragraph is to opt for a single-camera video and an imposed natural language question. Details of the proposed FRM, CRM, and MRM modules are described respectively in Subsections 4.1, 4.2 and 4.3.

Video feeds from multiple camera feeds are needed to be taken into account for a successful multi-camera reasoning task, which may be required for specific types of questions. For example, in case of a rapid movement of the target person across multiple camera viewpoints, the imposed question might ask a question that may not directly be answered by looking at one camera footage. The proposed method can equip an arbitrary (eight cameras in an ideal setting) number of video feeds to perform the proposed memory-driven multilevel reasoning operation. The output of the proposed memory-driven multilevel reasoning for a single-camera, i.e. the output of the MRM module for a single-camera video feed, is considered as an input to the MCRM module. The MCRM accepts multiple inputs, which are resultant of the memory-driven multilevel reasoning for the single-camera video feed. That is for the question involving multiple video feeds, each of the single-camera feeds are passed through FRM, memory-update, and MRM, and finally, the outputs of all the MRM modules of each of the cameras are passed through the MCRM. The output of the MCRM is used to predict the final answer word for the multi-camera VideoQA setting.

In a nutshell, the single-camera task of the SVideoQA dataset requires the involvement of the FRM and MRM modules. There is also a memory update operation in between FRM and MRM modules. Again, for the multi-camera VideoQA, the output of each of the single-camera VideoQA is passed through MCRM to predict the final answer word. Figures 6 and 7, show the detailed operation respectively for the single-camera and multi-camera VideoQA. Details of the proposed method are further elaborated in the following subsections.

4.1. Frame-relation Module (FRM)

The FRM module serves the purpose of considering each possible combination of frames (8-frames in our experiment) with the question sentence phrases. It is to be noted only one phrase is considered at a time. Also, the phrase chunk is passed through the LSTM and the output is combined with 8-frames in each possible combination. In Equation 2, p is the respective phrase that is considered with the combination of frames for the video. The term p is the output of the LSTM for the respective phrase. Again, all the possible combinations of p and the set of 8-frames is passed through g_θ. Here, g_θ is a multilayer perceptron (MLP).

$$m_p^f = \sum g_\theta([f_1, ..., f_8], p) \tag{2}$$

As seen in Figure 6, it is seen that phrase p is combined 8-frames, and the FRM makes every possible combination of the grouping of the respective pairwise combination of frame groups and the phrase p. In a subsequent operation, all the outputs of g_θ are summed and then passed to the memory update operation. As seen in the Figure 6, the output of FRM and CRM is combined, which is then forwarded to the memory update operation. Input to the memory update operation is the concatenation of m_p^f and m_p^c respectively from Equations 2 and 3. As visible in the figures, each of the phrases of the question sentence is considered separately throughout the proposed method, so if there are P number of phrases, then both the FRM and CRM will occur P number of times to perform the reasoning process. The benefit of the FRM is to be able o capture the context-specific combination of the frames of the respective video, which is also conditioned by the phrase i.e. the conceptual chunk of the question sentence. The phrase by phrase consideration technique allows the proposed model to better capture the evolving context of the question sentence over time. It is to be mentioned that the FRM and CRM are complementary operations that are necessary and required to successfully capture all the appearance and motion-based evolution of the context of the respective video.

4.2. Clip-relation Modle (CRM)

The motion-based features of the video may also be of interest, through the imposed open-ended natural language question. Thus, just frame-level feature-based reasoning will never be sufficient for an ideal VideoQA model. The proposed model considers both the frame and clip-level feature to better capture the evolving context of the video. As seen in Figure 6, the phrase p is given as input along with clip-level motion features to the CRM. Later, the CRM makes all possible combination of the respective phrase and clip-level motion features, which are then passed through the MLP g_θ. It is to be noted the MLP g_θ is different for FRM and CRM, termed respectively as g_θ^a and g_θ^m in the Figure 6.

$$m_p^c = \sum g_\theta([c_1, ..., c_8], p) \tag{3}$$

$$\mathbf{m^{in}} = LSTM(concat[m_p^f, m_p^c]) \tag{4}$$

The phrase-conditioned operation of FRM and CRM result in the outputs, which are concatenated together and passed through an LSTM as shown in Equation 4. The vector $\mathbf{m^{in}}$ is forwarded to the memory-update operation. In this manner, the phrase-conditioned motion and appearance-aware relation vector is forwarded to the memory update operations, which will later be followed by MRM.

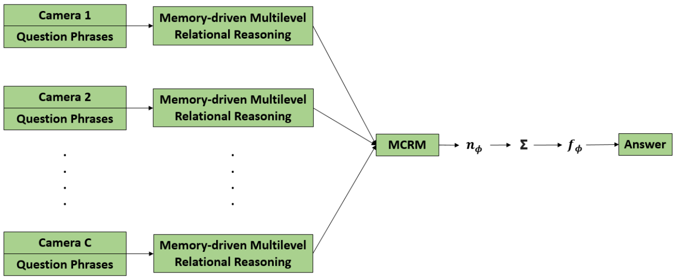

Camera 1
Question Phrases
Memory-driven Multilevel Relational Reasoning
Camera 2
Question Phrases
Memory-driven Multilevel Relational Reasoning
Camera C
Question Phrases
Memory-driven Multilevel Relational Reasoning
MCRM
n_ϕ
Σ
f_ϕ
Answer

Fig. 7. Each of the camera is considered separately for the memory-driven multilevel relational reasoning, if a question is imposed on the video feeds of multiple cameras, which is then passed individually to the MRCM. Each of the output of MRCM is forwarded through multilayer perceptron n_ϕ. All outputs of the n_ϕ are then summed together and forwarded to the multilayer perceptron f_ϕ to predict the final answer word.

4.3. Memory-relation Module (MRM) and Memory Update operation

The memory update operation is performed by the MRM module to extract the final logit vector for the answer word prediction. MRM produces all possible combinations of the vectors of the memory update operation and then passes those vectors through m_ϕ, which is an MLP.

$$\mathbf{m_p^{out}} = memory\_update(\mathbf{m^{in}}) \tag{5}$$

$$Answer, \mathbf{A} = f_\phi(\sum m_\phi(\mathbf{m_1^{out}}, ..., \mathbf{m_8^{out}})) \tag{6}$$

As seen in Equation 5, the LSTM output of the combination of the FRM and CRM is the vector m^{in}, which is derived from Equation 4. The vector m^{in} is passed through the memory update operation. Then, MRM considers all possible combinations of m_p^{out} in a grouping of eight different vectors, which are then combined and passed through the MLP f_ϕ. The output of f_ϕ leads to the final answer word. The memory update operation is described in the following paragraphs. The memory update operation described in [**?**] is adopted in the proposed method of this chapter.

$$c_t = \sigma(W_{oc}m^{in} + W_{hc}h_{t-1} + b_c) \tag{7}$$

$$a_t = v_a^T \tanh(W_{ca}c_t + W_{ha}h_{t-1} + b_a) \tag{8}$$

$$\alpha_{t,i} = \frac{\exp(a_{t,i})}{\sum_{j=1}^{S} \exp(a_{t,j})} \quad for, i = 1, 2, ..., S \tag{9}$$

$$e_t = v_e^T \tanh(W_{ce}c_t + W_{he}h_{t-1} + b_e) \tag{10}$$

$$\epsilon_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{3} \exp(e_{t,j})} \quad for, i = 1, 2, 3 \tag{11}$$

$$r_t = \sum_{i=1}^{S} \epsilon_{t,i} m_i \tag{12}$$

$$h_t^v = \sigma(W_{hh}^v h_{t-1}^v + W_{rh}^v r_t + b_h^v) \tag{13}$$

$$m_p^{out} = h_t^v \tag{14}$$

There is S number of slots in the memory with hidden state h considered for the memory update operation. The input to the memory update operation is the LSTM encoded combination of the output of FRM and CRM. The purpose of this memory update will be to write content in each of the S memory slots along with updating the hidden state h. The updated hidden state at the very end of the update operation will be considered as final memory output to be used in the subsequent operation of the proposed method. Contents to write in the memory are defined in Equation 7. The term c_t holds the non-linear mapping of the input to update the memory slots. The summation of all weights associated with each of the memory slots is a probability distribution, which results in 1 as shown in Equation 9. Also, the term a_t in Equation 8 is weighted to associate with a read operation; the weight is conditioned on the current and the previous hidden states of the memory LSTM.

The final hidden state of the memory is updated by following the read operation of the memory slots. Read weights are defined in Equation 10, which also sums to 1 as probability distributions of weights across S memory slots. Finally, the hidden state of memory is updated in Equation 13, where r_t is the weighted sum of each of the memory content. The output of the memory update operation is passed through the MRM as described at the beginning of Subsection 4.3.

5. MEMORY-DRIVEN MULTILEVEL RELATIONAL REASONING FOR SURVEILLANCE VIDEO QUESTION ANSWERING

5.1. Surveillance Video Question Answering

To adapt to the aspect of surveillance monitoring, multi-camera video feeds are included in the formulation with the existing formulation of the VideoQA problem as shown in Equation 15. In Equation 15, question Q is imposed on the several feeds of video, which include a feature set from F_1 to F_C, where C denotes the total number of cameras in the surveillance network.

$$Answer, \mathbf{A} = Model(\mathbf{Q}, F_1, ..., F_C) \tag{15}$$

5.2. Multi-camera Relation Module (MCRM)

The proposed multi-camera relation network adheres to several outputs of multiple cameras together, which are related to the imposed open-ended natural language question. To achieve this goal, each of the single-camera videos is first propagated through the memory-driven multilevel relational reasoning process as described in Section 4. The summation of MRM 6 for each of the single-camera is considered as input to the MCRM as shown in Figure 7. If there are C the number of cameras, then all possible combinations of the output of MRM for each of cameras are passed as input to the MCRM. Each of the MCRM output is then forwarded through n_ϕ, which is a multilayer perceptron. Finally, all the outputs from n_ϕ are summed together and forwarded to the f_ϕ multilayer perceptron to produce the final answer word.

In an ideal case, the proposed MCRM should be able to include any number of cameras, but with the proposed SVideoQA dataset, only a combination of up to eight cameras is considered as the upper limit. Open-ended questions are imposed on the overlapping regions of cameras as shown in Table 1.

$$Answer, \mathbf{A} = f_\phi(\sum n_\phi(\mathbf{c_1^{out}}, ..., \mathbf{c_8^{out}})) \tag{16}$$

6. EXPERIMENTS AND RESULTS

Experiments are performed on both the single-camera and the multi-camera open-ended questions. MSVD-QA and MSRVTT-QA datasets provide only single-camera open-ended questions. The proposed SVideoQA dataset provides both single-camera and multi-camera question on the surveillance video feeds. Subsection **??** provides the details of the proposed SVideoQA dataset. VideoQA task on the samples of MSRVTT-QA and the MSVD-QA dataset is shown in Figure 11. Also, an example of single-camera and multi-camera VideoQA task performed by the proposed method is depicted respectively in Figure 9 and Figure 8.

Single-camera Question-Answer Annotation	Multi-camera Question-Answer Annotation
Q: with which hand he touched his mouth? **A:** left **camera:** camera 6 **Q:** what is the color of his shirt? **A:** white **camera:** camera 7 **Q:** what is he holding in his hand? **A:** nothing **camera:** camera 1	**Q:** people are being seated on which side of his walking direction? **A:** left **region:** R167

Table 2. Open-ended question-answer pairs for single-camera and multi-camera video feeds of the SVideoQA dataset.

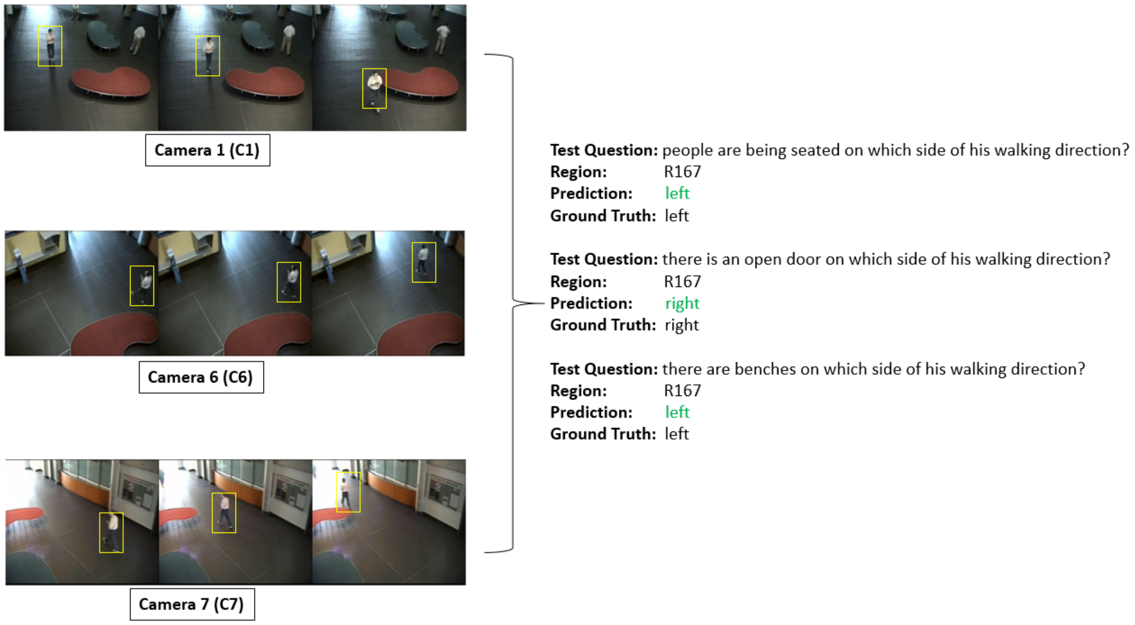
Methods	Accuracy on MSRVTT-QA	Accuracy on MSVD-QA
E-VQA of [3]	0.264	0.233
E-SA of [3]	0.293	0.276
E-MN of [3]	0.304	0.267
Gradual Attention [3]	0.325	0.320
heterogenour memory [?]	0.330	0.337
Hie. Rel. Attention [?]	0.3506	0.3439
multilevel relation	**0.390**	**0.393**

Table 3. Classes of question types.

6.1. Dataset

The proposed memory-driven multilevel relation network architecture is experimented with both the proposed SVideoQA dataset and other publicly available single-camera VideoQA dataset. Details about the SVideoQA dataset can be found in section 3.

MSVD-QA and MSRVTT-QA proposed in [3] are used with the proposed memory-driven multilevel relation network architecture. Both MSVD-QA and MSRVTT-QA provide open ended questions, where each of the question is associated with a single video. The MSVD-QA dataset is annotated on the videos of Microsoft Research Video Description Corpus dataset, which includes 1970 videos with 50505 open-ended questions. Again, the MSRVTT-QA dataset consist of 10,000 video clips with 243000 open-ended questions.

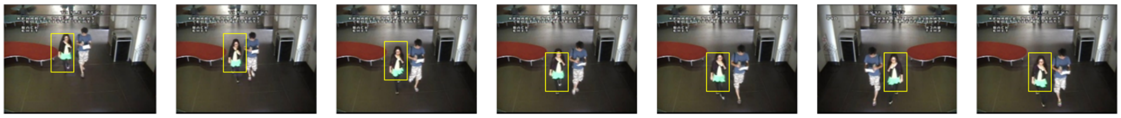


Test Question: people are being seated on which side of his walking direction?
Region: R167
Prediction: left
Ground Truth: left

Test Question: there is an open door on which side of his walking direction?
Region: R167
Prediction: right
Ground Truth: right

Test Question: there are benches on which side of his walking direction?
Region: R167
Prediction: left
Ground Truth: left

Fig. 8. The three rows of images respectively shows representative frames from the three overlapping cameras (Camera 1, Camera 6 and Camera 7) of the region R167. Imposed open-ended natural language questions and the predicted output of the model is shown on the right side of the image.



Dataset:	SVideoQA	Dataset:	SVideoQA
Question:	with which hand she is holding the cup?	Question:	with whom she is walking?
Prediction:	left	Prediction:	man
Ground Truth:	left	Ground Truth:	man

Fig. 9. Qualitative output on test questions for the single-camera task is shown in the image. The subject of the imposed question is marked in yellow bounding box. Single camera open-ended questions are annotated based on the target person.

Class of Question Type	Question Types
Action Search	Walking Direction
Group Activity Search	Counting Walking Companion
Attribute Search	Person Attribute Object Use

Table 4. Classes of different qustion types.

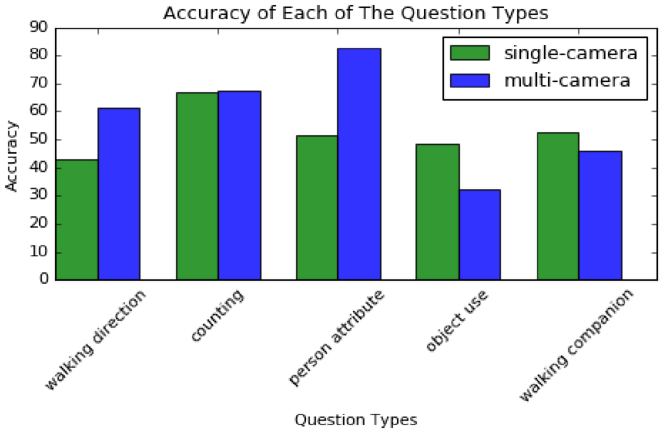


Fig. 10. Accuracy of the proposed model on each of the question types of the SVideoQA dataset.

6.2. Model Training Details

The MSVD-QA and MSRVTT-QA dataset are considered for the training of the single-camera SVideoQA task. The proposed SVideoQA dataset is used for both the single-camera and multi-camera VideoQA tasks. Appearance and motion-based features are extracted respectively by the pretrained VGG [13] and C3D [14] networks, which are trained respectively on the Sports-1M [15] dataset and ImageNet [13] dataset. In all cases, 16 consecutive frames of any video are considered as a clip to extract respective features of the video, and the dimensions of both the appearance and motion-based features is 4096. Features from the last fully connected layers are extracted for the model training purpose.

The embedding for each of the question sentence token is obtained through the GLOVE [16] embedding, which is pre-trained on Wikipedia 2014 [16] and Gigaword 5 [16] text repository. In exceptional cases where any word is out of the vocabulary, then the average of all other embeddings are considered for the particular token. The pretrained GLOVE embedding is 300 dimensional, and the size of the LSTM is also 300. Also, NLTK [**?**] is used to extract phrase chunks from the question sentences.

The single-camera task 6 uses the multilayer perceptrons g_θ, m_ϕ and f_ϕ. The g_θ and m_ϕ are multilayer perceptrons with 256 units, which are followed by the multilayer perceptron f_ϕ with the number of hidden units equivalent to the number of answers of the respective testing dataset. Again, in the multi-camera experimental settings, the multilayer perceptron n_ϕ holds 256 hidden units, which is followed by f_ϕ with the hidden unit number equivalents to the number of classification classes. The MSVD-QA and the MSRVTT-QA dataset are considered with a batch size respectively of 32 and 64. The SVideoQA dataset is considered with a batch size of 64. In all cases, a default learning rate of 3e-5 is chosen, and the cost function is minimized with the ADAM optimizer. A memory size of 30 is chosen with 256 hidden states for the memory update operation.

Dataset	Accuracy on Single-camera SVideoQA Task	Accuracy on Multi-camera SVideoQA Task
SVideoQA Dataset	0.570	0.576

Table 5. Accuracy of the proposed memory-driven multilevel relational reasoning model on the Single-camera and multi-camera VideoQA task of the SVideoQA dataset.

Question Types	Accuracy on Single-camera SVideoQA task	Accuracy on multi-camera SVideoQA task
Walking Direction	0.428	0.61217
Counting	0.667	0.676
Person Attribute	0.518	0.824
Object Use	0.483	0.325
Walking Companion	0.526	0.462

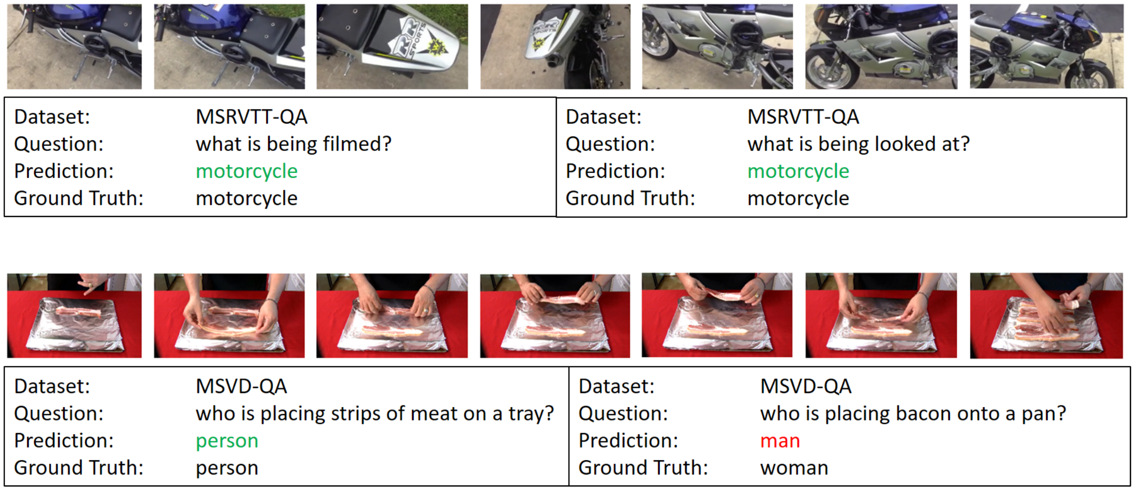
Table 6. Performance accuracy of the proposed memory-driven multilevel relational reasoning on each of the question types of the proposed SVideoQA dataset.

6.3. Qualitative and Quantitative Results

6.3.1. Results on SVideoQA

Open-ended questions impose the challenge of a real-time like reasoning requirement to the VideoQA models. The context of the imposed question is never known prior. The proposed memory-driven multilevel relation network architecture is tested for the VideoQA task with open-ended question-answer annotations. The proposed SVideoQA dataset consists of both single-camera and multi-camera open-ended question-answer annotations.

The Table 5 shows the performance of the proposed memory-driven multilevel relation network architecture on the proposed SVideoQA dataset. Figure 10 shows the accuracy of the proposed model on each of the question types of



Dataset: MSRVTT-QA
Question: what is being filmed?
Prediction: motorcycle
Ground Truth: motorcycle

Dataset: MSRVTT-QA
Question: what is being looked at?
Prediction: motorcycle
Ground Truth: motorcycle

Dataset: MSVD-QA
Question: who is placing strips of meat on a tray?
Prediction: person
Ground Truth: person

Dataset: MSVD-QA
Question: who is placing bacon onto a pan?
Prediction: man
Ground Truth: woman

Fig. 11. The MSRVTT-QA and MSVD-QA dataset provides open-ended natural language questions associated with single videos. First row and second row of the images show the qualitative output of the proposed model on the open-ended natural language questions respectively from the MSRVTT-QA and the MSVD-QA dataset.

the SVideoQA dataset. The *walking direction* type questions show higher accuracy in the case of the multi-camera setting of the proposed model. Accuracy single-camera setting shows higher accuracy in the case of predicting an answer for the *person attribute* type questions.

6.3.2. *Results on MSRVTT-QA and MSVD-QA*

Only single-camera video feeds along with the question-answer annotation pair are available in the MSVd-QA and MSRVTT-QA dataset. Table 6 provides the performance comparison between the accuracy of the baseline system and the proposed memory-driven multilevel relation network for the single-camera VideoQA task on MSVD-QA and MSRVTT-QA dataset. The proposed memory-driven multilevel relation network demonstrates a major performance improvement over the accuracy of the baseline methodology [3].

6.4. Grad-cam Visualization of the Single-camera and Multi-camera SVideoQA Task

Visualizing the focused portion of the visual features for the single-camera and multi-camera task is the best way to evaluate the performance variations for the single-camera and multi-camera SVideoQA task. Each of the question types is described below with the focus of the proposed model in the visual features. The focused visual portion describes why in some cases the proposed performs not very well even after availability of more data from multiple cameras. Grad-cam [**?**] is used to visualize the proposed model's focus on the relevant visual areas for the posed natural language questions.

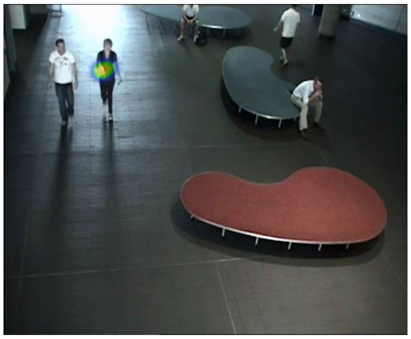


Fig. 12. The proposed model can successfully focus on the intended visual area for the question, ‘In which hand was she holding the documents?’ The target area is not occluded, and the model can also relate the target visual area aligned with the posed natural language question.

6.4.1. *‘Object Use’ Question Type*

The ‘Object Use’ type is one of the two question types where the proposed model fails to perform better even after the availability of more visual features across multiple cameras. Figure 12 and 13 refers to the scenario where an ‘Object Use’ type multi-camera question is posed, but the model fails to successfully predict the correct answer. The inherent cause of this variation in the performance is the occlusion of objects from the captured data. To elaborate more, let us consider the question ‘in which hand was she holding the documents?’ for the multi-camera task. As shown is Figure 12, in the case of ‘Camera 1’ the visibility, is evident with the network being able to successfully focus on the right hand of the target

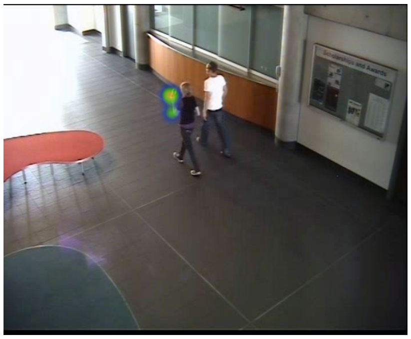


Fig. 13. The target area (right hand) is occluded due to the capturing angle and the physical orientation of the person. Thus for the question, ‘in which hand was she holding the documents?’, the proposed model fails to perform well even with more data being available from multiple cameras.

person, who hold the documents. On the other hand, as seen in Figure 13 for ‘Camera 7’, the visibility of the documents being held in the right is occluded due to the camera capture and the target person’s physical orientation.

The performance of the proposed model for the single-camera task for the ‘object use’ question type is 0.483. On the contrary, the accuracy of the same question type for the multi-camera task is 0.325. From the visualization of the proposed model’s focus on the visual aspect, it is evident that the underlying cause of the variation in performance is the occlusion of objects across multiple cameras. Even after the availability of the target, the questioned object is being occluded, and thus the model fails to successfully focus on it to answer the posed natural language question.

6.4.2. ‘Walking Companion’ Question Type

The other question type is ‘walking companion’ where the proposed model performs poorly even after the availability of more data from multiple-cameras. Most of the annotated questions for the ‘walking companion’ question types require the proposed model to determine the gender of the person accompanying the target person. From the grad-cam visualization of the proposed model, it is evident that the proposed model is not quite successful in determining the gender of the captured person. Also, the model struggles to determine who is accompanying whom when multiple persons are being captured across several cameras.

The capturing angle across multiple cameras is never front faced. Also, it is a normal practice to install surveillance cameras across a good height for a wider view for manual surveillance monitoring purposes. The physical orientation of the captured individuals makes it harder for the proposed model to determine the gender of the individual persons. Also, the illumination changes a potential cause that makes the proposed model perform worse in determining the gender across multiple cameras. Successful determination of the gender is the key for the ‘walking companion’ type questions because the ground truth answer is either ‘man’ or ‘woman’ in most of the cases.

6.4.3. Other Question Types

Except for the ‘object use’ and ‘walking companion’ question types described in previous sections, the availability of more data from multiple cameras aids the performance improvement for the multi-camera task. For example, let us consider the ‘walking direction’ type question as shown in Figure 15. The posed question is, ‘there are benches on which side of his walking direction?’. The availability of more data from multiple cameras improves the model’s stance in predicting the right answer. Also, the grad-cam visualization shows that the model can focus on the relevant visual features to successfully predict the right answer.

7. CONCLUSION AND FUTURE WORKS

The proposed memory-driven multilevel relation network model relies on finding relationship among various parts of video based on the context of the question through updating memory. The proposed model takes the advantage of conditioning the context of the relation network operation through considering each of the question phrase at once, and the proposed consideration is more like the human-like reasoning of any question sentence. Also, the aspect of including a multi-camera network video architecture with the proposed model makes the proposed model suitable for use in automated surveillance monitoring where multimodal inference of both video and text is required.

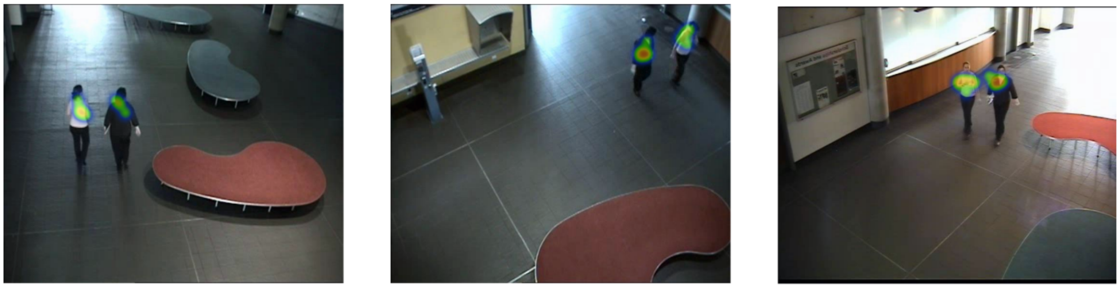
8. REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, “Vqa: Visual question answering,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[2] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun, “Leveraging video descriptions to learn video question answering.,” in *AAAI*, 2017, pp. 4334–4340.

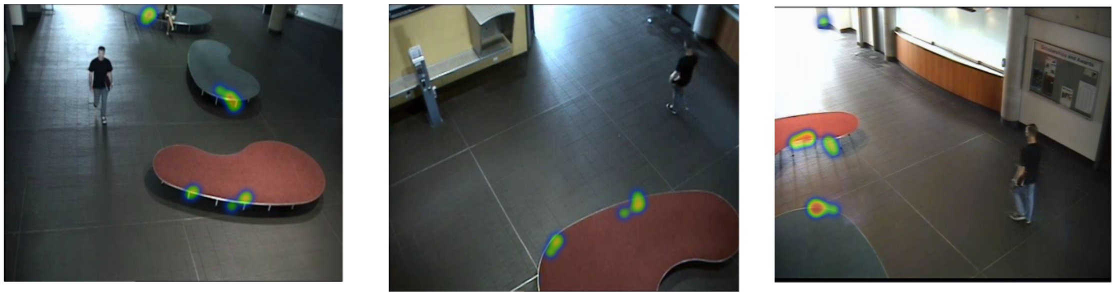
[3] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang, “Video question answering via gradually refined attention over appearance and motion,” in *ACM Multimedia*, 2017.

[4] Caiming Xiong, Stephen Merity, and Richard Socher, “Dynamic memory networks for visual and textual question answering,” in *International Conference on Machine Learning*, 2016, pp. 2397–2406.



Camera 1 Camera 6 Camera 7

Fig. 14. The proposed model finds it harder to determine the gender of the accompanying person due to the variance in physical orientation across multiple cameras. In this particular case, the posed question is 'with whom is she walking?', but the model fails to produce the right answer due to the variety in the physical orientation of the accompanying person.



Camera 1 Camera 6 Camera 7

Fig. 15. Availability of more data from multiple cameras helps the model to perform better in case of this 'walking direction' type question. The posed question is 'there are benches on which side of his walking direction?', and the model can successfully determine the correct answer by correctly focusing on the relevant regions in the visual features, as shown with the grad-cam visualization.

[5] Adam Santoro, David Raposo, David GT Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap, "A simple neural network module for relational reasoning," *arXiv preprint arXiv:1706.01427*, 2017.

[6] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1–9.

[7] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia, "Abc-cnn: An attention based convolutional neural network for visual question answering," *arXiv preprint arXiv:1511.05960*, 2015.

[8] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 21–29.

[9] Huijuan Xu and Kate Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *European Conference on Computer Vision*. Springer, 2016, pp. 451–466.

[10] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," *arXiv preprint arXiv:1704.04497*, 2017.

[11] Hongyang Xue, Zhou Zhao, and Deng Cai, "Unifying the video and question attentions for open-ended video question answering," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5656–5666, 2017.

[12] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang, "Video question answering via attribute-augmented attention network learning," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2017, pp. 829–832.

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

[14] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[15] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

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