Bi-Directional Attention Flow for Video Alignment

Reham Abobeah[1,2], Marwan Torki[3], Amin Shoukry[1,3] and Jiro Katto[4]
[1]*CSE Department, Egypt-Japan University of Science and Technology, Alexandria, Egypt*
[2]*CSE Department, Al-Azhar University, Cairo, Egypt*
[3]*CSE Department, Alexandria University, Alexandria, Egypt*
[4]*Computer Science and Communication Engineering Department, Waseda University, Tokyo 169-8555, Japan*

Keywords: Temporal Alignment, Synchronization, Attention Mechanisms, Bi-directional Attention.

Abstract: In this paper, a novel technique is introduced to address the video alignment task which is one of the hot topics in computer vision. Specifically, we aim at finding the best possible correspondences between two overlapping videos without the restrictions imposed by previous techniques. The novelty of this work is that the video alignment problem is solved by drawing an analogy between it and the machine comprehension (MC) task in natural language processing (NLP). Simply, MC seeks to give the best answer to a question about a given paragraph. In our work, one of the two videos is considered as a query, while the other as a context. First, a pre-trained CNN is used to obtain high-level features from the frames of both the query and context videos. Then, the bidirectional attention flow mechanism; that has achieved considerable success in MC; is used to compute the query-context interactions in order to find the best mapping between the two input videos. The proposed model has been trained using 10k of collected video pairs from "YouTube". The initial experimental results show that it is a promising solution for the video alignment task when compared to the state of the art techniques.

1 INTRODUCTION

Video Alignment aims at finding, in time and space, the best correspondences between two videos. More specifically, temporal alignment or synchronization refers to mapping each frame in a reference sequence to the most similar one in an input sequence, taking into consideration the sequence information (Diego et al., 2011). Along the last years, video alignment has played a significant role in many computer vision applications including video matting (Sand and Teller, 2004), action recognition (Ukrainitz and Irani, 2006), object detection (Kong et al., 2010), change detection (Diego et al., 2011), and video editing (Wang et al., 2014).

Most of the previous video alignment techniques have some restrictions on the captured videos such as constraining the used cameras to be either still or stereo or independently moving. Other difficult assumptions include knowing the trajectories of some feature points along the whole video (Singh et al., 2008) or supposing the existence of some linear relationship among the two videos (Padua et al., 2010). We think that these restrictions are among the reasons behind the limited applicability of these techniques in real applications. So, in this paper, we aim to overcome most of these limitations and open the way to new applications through our new proposed technique.

Recently, attention mechanisms have achieved great success in many applications in both NLP and computer vision areas (Weston et al., 2015; Agrawal et al., 2017; Xiong et al., 2016). Specifically, they improved in a significant way the performance of Recurrent Neural Network (RNN), through guiding it to "where to look" during the task. The most common characteristics of these attention-based works can be summarized in the following points. First, they extract the information related to the query through summarizing the context into a fixed length vector based on the attention weights. Second, the attention weights at any time step depend mainly on the attended vector from the previous step. Third, the attention direction is always from the query to the image, in captioning task or the context paragraph in NLP task.

Unlike these mechanisms, we depend, mainly in our work, on the bidirectional attention flow (BIDAF) technique introduced recently in (Seo et al., 2016). BIDAF outperforms the previous techniques by the

Abobeah, R., Torki, M., Shoukry, A. and Katto, J.
Bi-Directional Attention Flow for Video Alignment.
DOI: 10.5220/0007524505830589
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 583-589
ISBN: 978-989-758-354-4

following features which make it the best choice for our task. First, instead of context summarization, the attention is estimated at each time step and the output vector together with the previous layer representation are passed to the modeling layer. Second, the attention at each time step depends only on the query and the context at the current time step regardless of the attention at the previous step. Consequently, this prevents inaccurate attention at previous steps from affecting the current time attention. Finally, BIDAF technique applies the attention in both directions from the query to the context and vice versa. This, in turn, leads to a significant improvement in the overall accuracy through feeding more useful information to the model.

In this paper, we propose a new solution for the video alignment task. Specifically, we deal with the task as MC by letting one of the two input videos acts as a query and the other as the context, allowing the technique to find the best match for the given query. Nothing more than the existence of an overlapping among these two videos is required. We exploit all the previously mentioned features of the BIDAF mechanism and apply it to our task.

Through testing the proposed approach on a small challenging dataset, we conclude that it can be successfully applied to the state of the art video alignment datasets.

2 RELATED WORK

2.1 Video Alignment

There are many available solutions for the video synchronization task. They differ according to the restrictions and assumptions imposed by each method. We briefly introduce some of these previous works in this section.

Regarding the temporal correspondence, some works assume a constant offset time between each two corresponding frames $t_r = t_q + \beta$, (Tuytelaars and Van Gool, 2004; Ushizaki et al., 2006; Wolf and Zomet, 2006). Others, assume a linear relationship in order to consider the frame rate difference $t_r = \alpha\, t_q + \beta$, (Wedge et al., 2007; Tresadern and Reid, 2009; Ravichandran and Vidal, 2011). In both cases, the alignment models need to only estimate one or two parameters whereas in non-parametric curve assumption based models, the problem turns to be much harder (Sand and Teller, 2004; Fraundorfer et al., 2007; Ho and Newman, 2007; Singh et al., 2008; Cao et al., 2010).

By considering the relation between the used cameras, the existing solutions depend on using either two rigidly connected or independently moving cameras. The problem is easier to solve in the first case as the geometric transformation among the coordinate systems of both cameras is assumed to be constant (Wolf and Zomet, 2006; Wedge et al., 2007; Ravichandran and Vidal, 2011). On the contrary, no geometric relationship can be assumed in case of the two moving cameras. Consequently, the proposed solutions have to assume the existence of an overlapping field of view between the two cameras in order to solve the problem (Sand and Teller, 2004; Dai et al., 2006; Fraundorfer et al., 2007; Singh et al., 2008; Tresadern and Reid, 2009; Padua et al., 2010).

Also, the existing alignment models can be classified, according to their inputs, into two types, direct methods and feature based methods. In the first type, the model deals directly with the frame intensity values (Caspi and Irani, 2002; Ushizaki et al., 2006), its Fourier transform (Dai et al., 2006), or its dynamic texture (Ravichandran and Vidal, 2011). On the other hand, feature based methods require tracking the feature points along the whole two sequences (Tuytelaars and Van Gool, 2004; Wolf and Zomet, 2006; Singh et al., 2008), or both the feature points and lines along three sequences (Lei and Yang, 2006) or identifying the interest points in space or in space and time (Sand and Teller, 2004; Fraundorfer et al., 2007; Tresadern and Reid, 2009; Cao et al., 2010; Padua et al., 2010). It is worthy to note that some of the previous works can align the two videos even if they are captured at different points in time like (Ukrainitz and Irani, 2006; Ho and Newman, 2007; Singh et al., 2008; Cao et al., 2010), while other works deal only with simultaneously captured videos like (Caspi and Irani, 2002; Dai et al., 2006; Lei and Yang, 2006; Tresadern and Reid, 2009; Padua et al., 2010; Ravichandran and Vidal, 2011). Recently, (Douze et al., 2016) has introduced an alignment method that can accurately find the best correspondences among two overlapped videos even with a significant change in the view point. However, this method assumes a constant offset time among the two input videos.

Unlike the previous solutions, we aim to present a new alignment technique that works smoothly with no restrictions, other than the existence of an overlap among the two given videos.

2.2 Visual Question Answering (VQA)

Recently, the VQA task or answering a question about an image has been solved by representing both the question and the image through RNN and CNN, re-

spectively. Then, they are combined together to give a suitable answer for the question (Malinowski et al., 2015; Agrawal et al., 2017).

Also, some researches have exploited the attention mechanisms to solve the VQA task. They can be classied according to the granularity level of their attention mechanism and the attention matrix computation method. A coarse granularity level causes multiple image patches to gain attention from the question (Xiong et al., 2016; Zhu et al., 2016). While a finer level causes each question word to pay attention to all image patches and, at the end, the patch with the highest attention value is selected (Xu and Saenko, 2016). Also, (Yang et al., 2016) introduced the idea of combining the question representation at different levels of granularity (uni/bi/trigrams). On the other hand, there are various approaches to construct the attention matrix including concatenation, element-wise (sum/product) and Bilinear Pooling (Fukui et al., 2016).

Besides the work introduced by (Seo et al., 2016) in MC, that applied the bidirectional attention mechanism to their work, (Lu et al., 2016) also adopted the same idea. They proved that estimating the attention from the question words to the image patches and vice versa, has a great effect on the efficiency of the VQA task. Consequently, we expect the same success in solving the video alignment task using the same mechanism.

3 THE PROPOSED MODEL

Our alignment technique, as shown in Figure 1, consists of five layers that can be described as follows:

1. **Input Layer:** maps each frame from both query and context videos to a fixed length descriptor using a pre-trained CNN.
2. **Contextual Layer:** considers the relations between all the sequence frames in order to improve the feature extraction layer.
3. **Attention Layer:** ties the query and the context video vectors such that it generates a set of feature vectors representing the query-awareness degree for each frame in the context video.
4. **Modeling Layer:** is responsible for scanning the context video through applying the RNN.
5. **Output Layer:** gives the start and end indices for the best aligned part from the context video to the input query.

1. **Input Layer:** It is considered a frame level feature extraction layer. Let $\{x_1,...,x_M\}$ and $\{q_1,...,q_N\}$ be frames of a context and query videos, respectively. Each frame in the input is fully described by a fixed length vector through using a pre-trained VGG-16. VGG-16 is a convolutional neural network of 16 layers depth that is pre-trained on ImageNet dataset (Deng et al., 2009) which contains millions of static labeled images.

2. **Contextual Layer:** In this layer, a Long Short Term Memory Network (LSTM) (Hochreiter and Schmidhuber, 1997) is used in both forward and backward directions in order to consider the temporal interactions among input frames. The input to the contextual layer is the descriptors obtained from the first layer. Each LSTM output is of d-dimension and the final layer output is a concatenation of both LSTMs outputs to yield $Y \in R^{2d\times M}$ from the context video and $U \in R^{2d\times N}$ from the query video. Therefore, the dimension of each component of Y and U is 2d.

   It is worthy to mention that the first two layers of the model are computed for both the query and context and they represent two levels of feature extraction.

3. **Attention Layer:** This layer is in charge of fusing and linking the information from both the query and the context video frames. As mentioned earlier, it adopts a different way in computing the attention, other than those in the previous mechanisms. Instead of summarizing in a single representative vector both the query and the context, the attention of the query to each frame in the context video is estimated at each time step. This layer takes as input the contextual representation vectors of both the query and the context videos and outputs the query-aware representation vectors G of the context video. In addition, the context layer output is passed to the Modeling layer. One of the main pros of this layer is computing the attention in both directions to obtain, at the end, a set of highly representative attended vectors in each direction.

   The shared step in computing the bidirectional attention is to calculate a similarity matrix between both the context and the query frames as follows:

   Similarity Matrix: It is constructed by computing the cosine similarity between the m-th context frame and the n-th query frame to obtain, at the end, $S \in R^{M\times N}$.

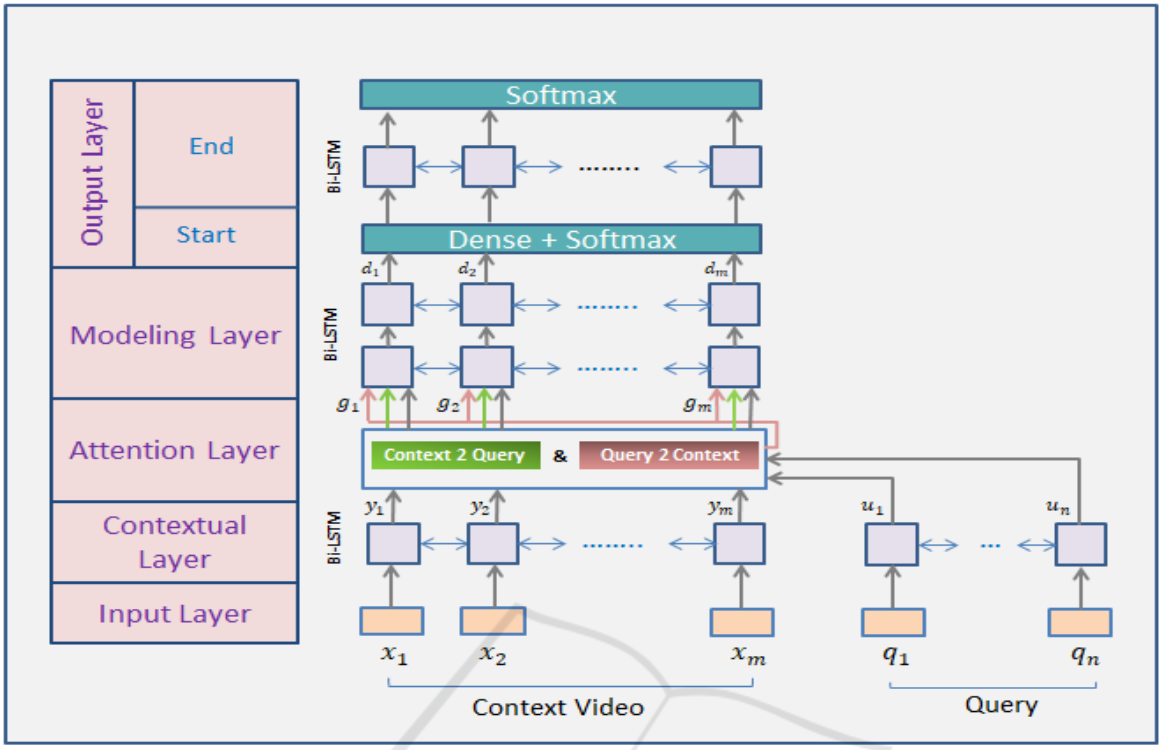

Figure 1: The BIDAF-based Video Alignment Model.

Context to Query Attention: This step determines the most relevant query frames to each frame in the context video. Given the similarity matrix S, the attention weights of the query frames to the m-th frame in the context video, are calculated according to the following formula:

$$a_{:m} = softmax(S_{:m}) \in R^N \quad (1)$$

where $S_{:m}$ is the m-th column vector in S and $\sum a_{mn} = 1$, for all m. Consequently, we obtain the $\bar{U}$matrix of size 2d by M, such that each column represents the attended query vector to a specific frame in the context video:

$$\bar{U}_{:m} = \sum_n a_{mn} U_{:n} \quad (2)$$

Query to Context Attention: This step is very critical for reaching the best answer to the input query. It determines the context video frame that has the best similarity to one of the query frames. First, the attention weights of each context frame are calculated using the following formula:

$$b = softmax(\max(S)) \in R^M \quad (3)$$

where max function is applied to each column in the S matrix. Then, the weighted context vectors are obtained using

$$\bar{Y}_{:m} = b_m Y_{:m} \quad \in R^{2d \times M} \quad (4)$$

Finally, the targeted attended vector is the weighted sum of these vectors which represent the most important context frames w.r.t. the query frames as follows:

$$\bar{y} = \sum_m \bar{Y}_{:m} \quad \in R^{2d} \quad (5)$$

The final output of this layer is obtained by combining the output of both the contextual and the attention layers. Semantically, each column in the output matrix is aware of each context frame representation and is calculated through

$$G_{:m} = [Y_{:m}; \bar{U}_{:m}; Y_{:m} \circ \bar{U}_{:m}; Y_{:m} \circ \bar{Y}_{:m}] \in R^{8d \times M} \quad (6)$$

where $\circ$ is an element-wise multiplication and $[\,;]$ is a concatenation of vector across row.

4. **Modeling Layer:** Unlike the contextual layer that computes the interactions among the context frames with no regard to the query frames, this layer captures these interactions by considering the query frames. It takes as input the output of the previous layer, G matrix, which reflects the query-awareness of each context frame. The modeling layer processes the input through two bi-directional LSTM layers, each direction output

Figure 2: Example of BIDAF-based Indoor Alignment. The top and the middle rows represent some selected frames from the query sequence and their corresponding frames in the original one, respectively. The bottom row represents the fusion image where the Red & Blue channels in the RGB image are assigned to the query frame, while the Green channel is assigned to its corresponding frame.

Figure 3: Example of BIDAF-based Outdoor Alignment. The top and the middle rows represent some selected frames from the query sequence and their corresponding frames in the original one, respectively. The bottom row represents the fusion image where the Red & Blue channels in the RGB image are assigned to the query frame, while the Green channel is assigned to its corresponding frame.

is of size "d", hence the final output is a matrix $D \in R^{2d \times M}$. Each column in the output matrix captures the contextual information of the m-th frame w.r.t the whole context and the query videos.

5. **Output Layer:** This layer aims at finding the best part in the context video corresponding to the input query. Specifically, it responds by the start and end of the frame indices of the best sub-sequence in the context video. Firstly, the probability distribution of the start index w.r.t. the whole video can be obtained through:

$$P_{Start} = softmax(\ w^T_{p_{start}}[\ G;D]\ ) \qquad (7)$$

Where $w_{p_{start}} \in R^{10d}$ represents a trainable weight vector. For obtaining the end frame index, the D matrix is pushed to one bi-directional LSTM layer in order to get $D^\star \in R^{2d \times M}$. Then, the probability distribution of the end index is estimated through:

$$P_{End} = softmax(\ w^T_{p_{end}}[\ G;D^\star]\ ) \qquad (8)$$

4 THE PROPOSED APPROACH EVALUATION

We train our model on 10k of YouTube videos that are aligned manually in an accurate way. For each query video, its corresponding part in the context video is identified by the start and end of frames indices.

As an initial evaluation, the proposed model is tested by the blind navigation dataset introduced in our recently published work (Abobeah et al., 2018). Although this dataset has been collected in the context of a navigation technique for blind people, it has a tight relation to our present task. The navigation technique depends mainly on identifying the current location of the blind user w.r.t. a reference video through using an on-line alignment technique. This dataset consists of 12 video pairs of 22 minutes total length, half of them are captured at outdoor and the rest at indoor. All videos are captured using a chest mounted mobile phone camera carried by a blind person along some indoor/outdoor routine paths and

are manually annotated for the instruction generation task in the navigation work. For our task, all videos are re-annotated such that each query, which is randomly chosen with various lengths from 5 to 120 seconds from each video, is aligned to its original video. Specifically, we assign the interval $[\ l_t, u_t]$ in the original video to represent the start and end indices of the query input, and given the predicted interval from the proposed model, we can estimate the alignment error for each pair.

We can qualitatively observe the performance of our proposed approach on two video pairs from the navigation dataset, one is captured at indoor as shown in figure 2, and the other at outdoor as shown in figure 3. Clearly, We can observe the well defined correspondences among each of the two videos, which support our idea of considering our approach a promising solution for the video alignment task.

5 CONCLUSIONS

In this work, we present a new technique to solve the temporal alignment task between two overlapped videos with no restrictions imposed on the capturing process. The proposed technique uses the pretrained CNN network,"VGG-16", to obtain highly descriptive features of the video frames. Also, it exploits the bi-directional attention flow mechanism that has already proved its efficiency in MC in order to consider the existing interactions between the two input videos in both directions. Initial results obtained using a training dataset including around 10k of video pairs from "YouTube" show that this approach is highly effective in mapping the input query video to its corresponding part in the context video. We plan to test our model using the state of the art datasets for video alignment to be able to assess its accuracy thoroughly.

ACKNOWLEDGEMENTS

This work has been supported by the Ministry of Higher Education (MoHE) of Egypt and Waseda University at Japan through a PhD scholarship.

REFERENCES

Abobeah, R., Hussein, M., Abdelwahab, M., and Shoukry, A. (2018). Wearable rgb camera-based navigation system for the visually impaired. In *Advanced Concepts for Intelligent Vision Systems*, volume 5, pages 555–562.

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., and Batra, D. (2017). Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31.

Cao, X., Wu, L., Xiao, J., Foroosh, H., Zhu, J., and Li, X. (2010). Video synchronization and its application to object transfer. *Image and Vision Computing*, 28(1):92–100.

Caspi, Y. and Irani, M. (2002). Spatio-temporal alignment of sequences. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1409–1424.

Dai, C., Zheng, Y., and Li, X. (2006). Accurate video alignment using phase correlation. *IEEE Signal Processing Letters*, 13(12):737–740.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee.

Diego, F., Ponsa, D., Serrat, J., and López, A. M. (2011). Video alignment for change detection. *IEEE Transactions on Image Processing*, 20(7):1858–1869.

Douze, M., Revaud, J., Verbeek, J., Jégou, H., and Schmid, C. (2016). Circulant temporal encoding for video retrieval and temporal alignment. *International Journal of Computer Vision*, 119(3):291–306.

Fraundorfer, F., Engels, C., and Nistér, D. (2007). Topological mapping, localization and navigation using image collections. In *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pages 3872–3877. IEEE.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Ho, K. L. and Newman, P. (2007). Detecting loop closure with scene sequences. *International Journal of Computer Vision*, 74(3):261–286.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kong, H., Audibert, J.-Y., and Ponce, J. (2010). Detecting abandoned objects with a moving camera. *IEEE Transactions on Image Processing*, 19(8):2201–2210.

Lei, C. and Yang, Y.-H. (2006). Tri-focal tensor-based multiple video synchronization with subframe optimization. *IEEE Transactions on Image Processing*, 15(9):2473–2480.

Lu, J., Yang, J., Batra, D., and Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297.

Malinowski, M., Rohrbach, M., and Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9.

Padua, F., Carceroni, R., Santos, G., and Kutulakos, K. (2010). Linear sequence-to-sequence alignment.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(2):304–320.

Ravichandran, A. and Vidal, R. (2011). Video registration using dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):158–171.

Sand, P. and Teller, S. (2004). Video matching. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 592–599. ACM.

Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Singh, M., Cheng, I., Mandal, M., and Basu, A. (2008). Optimization of symmetric transfer error for sub-frame video synchronization. In *European Conference on Computer Vision*, pages 554–567. Springer.

Tresadern, P. A. and Reid, I. D. (2009). Video synchronization from human motion using rank constraints. *Computer Vision and Image Understanding*, 113(8):891–906.

Tuytelaars, T. and Van Gool, L. (2004). Synchronizing video sequences. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.

Ukrainitz, Y. and Irani, M. (2006). Aligning sequences and actions by maximizing space-time correlations. *Computer Vision–ECCV 2006*, pages 538–550.

Ushizaki, M., Okatani, T., and Deguchi, K. (2006). Video synchronization based on co-occurrence of appearance changes in video sequences. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 71–74. IEEE.

Wang, O., Schroers, C., Zimmer, H., Gross, M., and Sorkine-Hornung, A. (2014). Videosnapping: Interactive synchronization of multiple videos. *ACM Transactions on Graphics (TOG)*, 33(4):77.

Wedge, D., Huynh, D., and Kovesi, P. (2007). Using space-time interest points for video sequence synchronization.

Weston, J., Chopra, S., and Bordes, A. (2015). Memory networks. *arXiv preprint arXiv:1410.3916*.

Wolf, L. and Zomet, A. (2006). Wide baseline matching between unsynchronized video sequences. *International Journal of Computer Vision*, 68(1):43–52.

Xiong, C., Merity, S., and Socher, R. (2016). Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406.

Xu, H. and Saenko, K. (2016). Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer.

Yang, Z., He, X., Gao, J., Deng, L., and Smola, A. (2016). Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29.

Zhu, Y., Groth, O., Bernstein, M., and Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004.