


RESEARCH ARTICLE

Open Access



# Lifelogging caption generation via fourth-person vision in a human–robot symbiotic environment

Kazuto Nakashima<sup>1\*</sup> , Yumi Iwashita<sup>2</sup> and Ryo Kurazume<sup>3</sup>

## Abstract

Automatic analysis of our daily lives and activities through a first-person lifelog camera provides us with opportunities to improve our life rhythms or to support our limited visual memories. Notably, to express the visual experiences, the task of generating captions from first-person lifelog images has been actively studied in recent years. First-person images involve scenes approximating what users actually see; therein, the visual cues are not enough to express the user's context since the images are limited by his/her intention. Our challenge is to generate lifelog captions using a meta-perspective called “fourth-person vision”. The “fourth-person vision” is a novel concept which complementary exploits the visual information from the first-, second-, and third-person perspectives. First, we assume human–robot symbiotic scenarios that provide a second-person perspective from the camera mounted on the robot and a third-person perspective from the camera fixed in the symbiotic room. To validate our approach in this scenario, we collect perspective-aware lifelog videos and corresponding caption annotations. Subsequently, we propose a multi-perspective image captioning model composed of an image-wise salient region encoder, an attention module that adaptively fuses the salient regions, and a caption decoder that generates scene descriptions. We demonstrate that our proposed model based on the fourth-person concept can greatly improve the captioning performance against single- and double-perspective models.

**Keywords:** Visual lifelogging, Image captioning, Intelligent space

## Introduction

Motivated by the availability of consumer wearable devices, lifelogging have been attracting increasing attention. By simply attaching a wearable device on their bodies, people can easily accumulate daily records of their states, activities, or experiences as *lifelog* data. Accumulated data are then analyzed and organized as an indexed digital collection that people can access whenever they want to review their lifestyle. For example, people with wrist-mounted sensors such as an Apple Watch can measure the number of steps, heart rates, or multi-axis

acceleration in order to analyze their activities. Such biometric data are widely used to estimate health level, stress level, and number of calories burned, but these applications are limited to quantification of ones' internal dynamics.

In contrast, lifelogging with images taken from a wearable camera such as SenseCam [1], GoPro HERO, and Narrative Clip offers us high-fidelity records of everyday visual experiences, which is specially referred to as *visual* lifelogging [2]. Wearable cameras are generally placed on the wearer's chest or head to get a first-person perspective such that the images involve everyday scenes showing what the wearer gazes, reacts to, manipulates, and any other interactions, throughout the day. By applying various methods of parsing image content, the pooled images are then tagged with characteristic attributes such as wearer's activities, objects, colors, among others, and

\*Correspondence: k\_nakashima@irvs.ait.kyushu-u.ac.jp

<sup>1</sup> Graduate School of Information Science and Electrical Engineering, Kyushu University, 744 Motooka Nishi-ku, Fukuoka-shi, Fukuoka 819-0395, Japan

Full list of author information is available at the end of the article

that allows the users to explore the collection by using keywords. In comparison to a non-visual lifelog, observing the wearer's social activities is easy in this case. Therefore, the first-person vision has traditionally been utilized not only for visual lifelogging purpose but also for social modeling and path prediction.

In visual lifelogging, to make it easy to explore the large collection, the acquired images are cleansed and structured with semantic tags that represent wearer's visual experiences. For example, at first, the uninformative images are filtered out [3] and the remaining ones are divided into homogeneous temporal segments [4]. Then, they are automatically indexed with predefined semantic tags such as the types of wearer's actions [5], places [6], and objects manipulated by hands [7], so that the user can search and retrieve the intended images/videos by specifying the visual characteristics in queries. The semantic tags can be extracted via various image recognition techniques such as object recognition, object detection, and semantic segmentation, which have been improving rapidly with the use of deep neural networks.

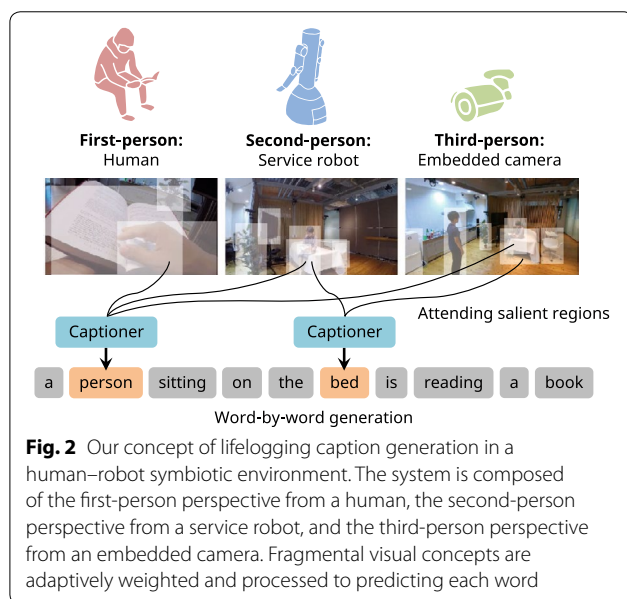
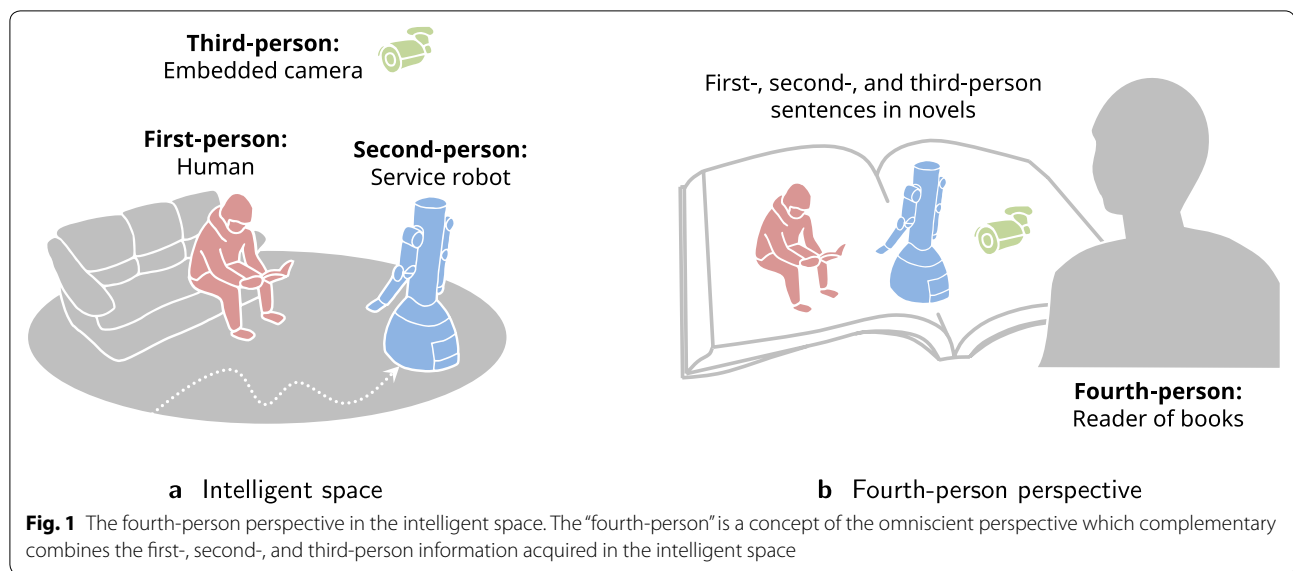
Till date, in the typical first-person vision field, recognition/detection of sports action and detection of the grasping objects, among others, have been designated as principal tasks. In the meantime, with the progress of recent deep learning techniques, the attempt to describe the first-person images with natural language sentences is advancing [8, 9]. Natural language descriptions not only simply list the visual concepts present in an image independently for each object but also represent their relationships in a natural and free form. Simultaneously, research on encoding and semantic understanding of natural language sentences is also progressing. Thus, the visual lifelogging field can evolve from a keyword-based system to a human-friendly, text-based system. That offers us an accessible interface for visual lifelogging; for example, instead of listing keywords like {"dog", "couch", "playing"} from a predefined word dictionary, you just have to command "the moment of playing with my dog on the couch" in a natural manner.

As mentioned above, conventional visual lifelogging has relied on the first-person wearable cameras directly capturing the wearer's visual experiences and object manipulation history. However, their visual information tends to be noisy or often has meaningless frames due to the wearer's dynamic ego-motion, occlusion by hands, and unintentional fixing at a wall and a ceiling, which may obscure the events of interest. To tackle this problem, many studies have proposed preprocessing such as keyframe detection to filter out such frames [2]. However, even if the camera succeeds in photographing the static scenes, it is still insufficient to understand the context of

the wearer's behavior from the limited beforehand scenes and the recorded collection is biased to static activities.

Therefore, herein, we consider combining the complementary observer's viewpoints. In this paper, we assume the multi-perspective vision system in the "intelligent space" wherein a human and a service robot coexist. The intelligent space is the room or the area that is equipped with various sensors or cameras, which has been widely studied in the robotics community because of its feasibility with regard to human-robot coexistence [10, 11]. Although it is difficult for a standalone robot to observe the dynamic environment and operate diverse service tasks for humans with only onboard sensors, an intelligent space enables it to expand its observation area. In this system, we can use a robot-view camera, not just the user's first-person viewpoint. The camera is movable to follow and capture human behaviors and interactions with a human closely. We define such a *camera agent* as the second-person viewpoint. Moreover, the typical intelligent space has embedded cameras on the wall or the ceiling to observe the comprehensive state, which is used to track the human and the robot. We define this type of camera as the third-person viewpoint. Those observer's viewpoints have the capability to capture exocentric information such as the user's postures and place types, which are important cues to complement the first-person description.

In this paper, we introduce the novel lifelogging concept "fourth-person vision", which complementary exploits the first-, second-, and third-person images as described above to generate accurate and detailed descriptions. The perspective term "fourth-person" was initially introduced into our previous study [12], which is an analogy of a storyteller or a book reader who picks up unique information within multi-perspective sentences and appreciates the storylines, as illustrated in Fig. 1. This paper aims to demonstrate that the fourth-person vision system improves the accuracy in the caption generation task required in the text-based visual lifelogging. To build this concept on the task, it is required to handle the visual complementarity and redundancy of the multi-perspective images and to learn visual-semantic relations, as depicted in Fig. 2. Therefore, we newly design a neural architecture to form a single natural language sentence describing the scene events from the synchronized multi-perspective images based on our concept. Through general caption evaluation schemes, we demonstrate that our proposed method can accurately generate sentences that contain visual attributes of multi-level granularity, as compared to methods with single or double input images. To the best of our knowledge, this is the first work that focuses



on multi-perspective images for improving caption generation. Our contributions can be summarized as follows.

- We propose a novel architecture to generate a sentence from multi-perspective lifelog images capturing the same moments in a human-robot symbiotic environment.
- We construct a new dataset composed of synchronized multi-perspective image sequences that are annotated with natural language descriptions for each sequences.

- We conduct experiments of caption generation on perspective ablation settings and demonstrate that our approach achieves significant improvements on common metrics in the image caption generation task.

## Related work

Our work relates to visual lifelogging and image captioning. In this section, we review related work on these topics and describe the approach of this study.

## Visual lifelogging

The typical procedures of the visual lifelogging system mainly consist of aggregating first-person images streamed from the user's wearable camera, filtering frames, and tagging them automatically. The position of the wearable camera is generally chosen from the user's head or chest [2]. The head-mounted camera provides the field of view that the wearer is possibly looking at and that can be used for modeling the wearer's attention, however, the acquired images are often noisy and blurred due to the ego-motion. In contrast, the chest-mounted camera has been widely used to record more stable and focused views of the wearer's manipulative workspace. Such visual information from the first-person perspective provides important clues to characterize the streamed images [5–7] and transform them into the indexed collection, so that the users can retrieve the image by specifying query keywords recalled by scenes. However, the motion of the first-person perspective is dynamic and un-intensional. Although the motion information itself contributes to wearer's action recognition task [13], the

acquired images often contain motion blur or meaningless frames such as ones filled with sky, ceiling, etc. A general approach to mitigating this problem is to detect and filter out the meaningless frames [3].

Most studies in visual lifelogging have been using the first-person wearable camera since its perspective is very close to camera wearers' visual experiences. In contrast, most visual recognition tasks reside on the observer's perspective because a large number of stable snapshot images can be aggregated from the web, which is one of the factors of recent advances in deep neural networks. In the context of visual lifelogging, such third-person perspective can be utilized through cameras fixed in the room [14]. The images are well-focused and less blur, however, a resolution for each object region and occlusions could be problems depending on the positioning. To take complementary advantages in the first and third-person perspectives, Sigurdsson et al. [15] proposed a neural mechanism to share knowledge between them for the human action recognition task. Xu et al. [16] simultaneously segmented identical person regions. Fan et al. [17] identified first-person images from multiple camera wearers in the third-person videos. On the other hand, our goal is to generate a caption describing the human and the context by jointly adopting the wearer's first-person perspective and the observer's perspective images. Particularly, we investigate the effects of second-person observer and third-person observer as mentioned in Sect. "Introduction".

### Image captioning

Beyond the task of recognizing the object categories and the spatial locations on images, recent studies include understanding the relationships between the objects. Particularly, image captioning has attracted considerable attention in the communities of computer vision and natural language processing [18–20]. The task is to generate captions describing a given image with respect to positional relationships or interaction between salient objects. Recent studies can be categorized into two main streams: dataset proposals to generate custom captions such as with styles and sentiment, and model proposals to improve caption quality on public benchmarks.

In terms of visual lifelogging, expressing an image in natural language is promising for text-based interfaces, and it has the potential to jointly describe social interactions, scene types, and human actions, which have been studied as separate tasks till date. As an initial study in this area, Fan et al. [8] applied existing image captioning technique [18] on the first-person images and evaluated the caption-based image retrieval system. Bolaños et al. [9] proposed a technique to generate captions conditioned by temporally ordered images.

### Datasets

Owing to a large number of images, the Microsoft COCO dataset [21] is widely used in image captioning benchmarking. COCO contains over 160k images, each with 5 reference captions, which were selected from Flickr "to gather images containing multiple objects in their natural context" [21]. There is no annotation protocol about the camera perspective or the photographers. Similarly, Flickr8k and the extended version Flickr30k also contain 5 captions for each image.

As another paradigm toward visual lifelogging application, Fan et al. [8] collected 696 first-person images obtained from chest-mounted wearable cameras (Narrative Clip) and produced an average of 14.7 captions per image. The sentences are built with two manners from a *grammatical* perspective. One describes the image through the third-person perspective to represent what is happening. The other is through a first-person perspective that instead focuses on the relationship between the camera wearer and the scene. For example, the first-person caption is diary-like sentence starting with "I am".

In contrast to the previous studies, we construct the caption dataset composed of multi-perspective images to generate lifelog captions. The experiments by Fan et al. [8] demonstrated that the COCO-trained model was not the best for evaluating their original dataset and one of the reasons was the difference in sentence styles. Therefore, we do not consider the grammatical perspective of sentences but unify them into the third-person style to utilize the visual diversity and word vocabulary of publicly available large datasets such as COCO. The collection procedure and the detailed statistics are described in Sect. "Dataset".

### Models

In general formulation, the captioning model is trained by maximizing the cumulative likelihood of words in a reference caption.

**ShowTell:** Vinyals et al. [18] analogized the automatic generation of image captions as "machine translation from an image to a description" and succeeded to train deep networks that generate a template-free caption from a given image. The approach extended the encoder–decoder model proposed in the machine translation field wherein an image feature is abstracted by a pretrained CNN encoder and the sequential likelihoods of vocabulary words are predicted step by step from an RNN decoder. Model training is formulated to maximize the cumulative log-likelihood of reference captions.

**ShowAttendTell:** In addition, Xu et al. [19] applied an attention mechanism that was proposed in machine translation field to improve the word alignments between the source and target languages. The attention

mechanism is a type of a dictionary model that has a set of feature candidates from an encoder. At each decoding step, the input features are adaptively selected with hard/soft weights computed by the top-down signal from a decoder. In this manner, the model can efficiently propagate the source information to predict the sequential results. In the context of the image captioning task, the attention mechanism receives the CNN feature maps as grid features so that it bridges specific image regions and prediction of each word in a caption. This approach has significantly improved caption quality.

**UpDown:** More recently, Anderson et al. [20] proposed a novel approach where the attention mechanism receives a set of region-of-interest (ROI) features as candidates, instead of the grid features as in *ShowAttendTell*. With this modeling, the visual concepts in both foreground and background appear in the image are encoded as object features while the *ShowAttendTell* model possibly disassembles them into several grid features. In this manner, the context and the relationships of salient objects can be accurately reflected in the caption.

Contrary to the *ShowTell* model that takes an image as a global feature to be used in the caption decoder, the attention-based models, *ShowAttendTell* and *UpDown*, can pool multiple feature candidates that are spatial grids or salient regions. Our key idea for multi-perspective image captioning is “attend to fuse”, that is, to organize the feature candidates across complementary multiple images acquired from a human–robot symbiotic environment. Moreover, we assume that the attention module pretrained with captions per image using a dataset such as COCO [21] generalizes to the feature candidates scattered on images that have different perspectives.

#### Fourth-person vision

In this section, we introduce a novel lifelogging concept, “fourth-person vision”. We assume the multi-perspective vision system in the “intelligent space” wherein a human and a service robot coexist. The intelligent space is the room or the area that is equipped with various sensors or cameras, which has been widely studied in the robotics community because of its feasibility with regard to human-robot coexistence [10, 11]. Herein, let us classify the available cameras in terms of the perspective.

##### First-person vision

As the typical source for visual lifelogging, we can use the images from the human’s wearable camera, which can be referred to as the first-person perspective. The first-person images involve objects that the user pays attention to or scenes of object-hand interaction, which are important clues to characterize the streamed images. However, their

visual information has problems such as lack of user context, noises, and occlusions, due to the unintentional self-motion and the narrow field-of-view.

##### Second-person vision

The co-living service robot can offer a different perspective. Unlike the other unintentional or fixed cameras, the service robot can control its viewpoint by moving so that it can also track and capture the dynamic human movements exclusively. Therefore, the viewpoint can produce stable visual cues for describing human activities. In addition, the user’s subtle behaviors can be observed up close in the interaction situations such as handing-over of household items and conversation. This perspective can be defined as the second-person.

##### Third-person vision

Last but not least, the environment has embedded cameras that are used to monitor and track the human and the robot. The embedded cameras offer the images comprehensively including the human, the robot, household items, and their interactions, which can be defined as the third-person perspective. However, the third-person cameras are not suitable to analyze detailed parts within the images such as human’s activities, since the cameras are fixed on the wall or ceiling and involve occlusions. To deal with the problems naively, it is required to increase the number of third-person cameras to fill up the gaps.

##### Fourth-person vision

Thus we complementary combine three types of perspectives acquired from the intelligent space to describe the scenes accurately. Notably, we refer the concept as to “fourth-person vision”. The perspective term “fourth-person” was initially introduced into our previous study [12], which is an analogy of storytellers or readers of books. As illustrated in Fig. 1b, the reader can pick up all intrinsic and extrinsic information from the first-, second-, and third-person sentences and build the omniscient perspective to understand the comprehensive situations accurately. In our daily life, the intrinsic information such as ones’ interests and sights is unknown to others. The proposed fourth-person aims to observe the information through the distributed cameras in the intelligent space and build the meta-perspective that complementary combines them to suppress the ambiguity in the scene description.

#### Our approach

In this paper, the task is defined to generate a single caption by jointly processing his/her first-person image, the second-person image from the service robot, and the third-person image from the embedded camera. To



configure such multi-image captioning model and to validate how much each perspective contributes to the resulting caption, we extend the state-of-the-art image captioning model UpDown by Anderson et al. [20]. The architecture is illustrated in Fig. 3. UpDown first enumerates salient regions within a given image, encodes the spatial feature into a fixed-size vector per region Sect. (“Image encoding”), and feeds them into the captioning process with an attention mechanism Sect. (“Caption generation”). For our multi-perspective situation, the region features are given from each perspective and are fed into the captioning process as attention candidates to decode words. In this study, we especially focus on how we can reorganize the attention candidates from the multi-perspective images. We propose a bottom-up fusion step that clusters the salient region features to suppress the appearance of the identical instances over multiple viewpoints Sect. (“Salient region clustering”).

### Image encoding

Suppose that we are given a set of synchronized multi-perspective images from the human’s egocentric viewpoint (first-person), bystander robot’s viewpoint (second-person), and birds eye viewpoint on a fixed camera (third-person). For each perspective, a set of region-of-interests (ROIs) are detected and encoded into the feature vectors according to their visual attributes, which

are attention candidates for subsequent captioning modules. We refer to the feature set of three images as  $V$ .

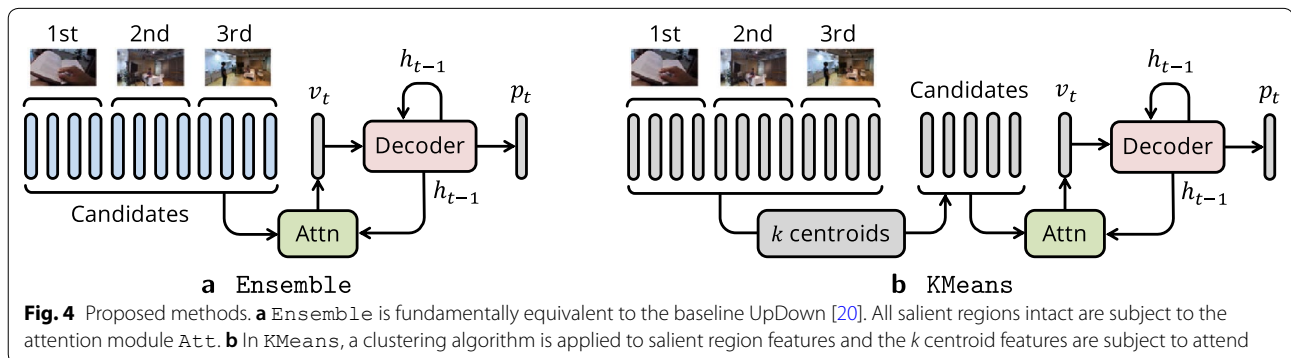
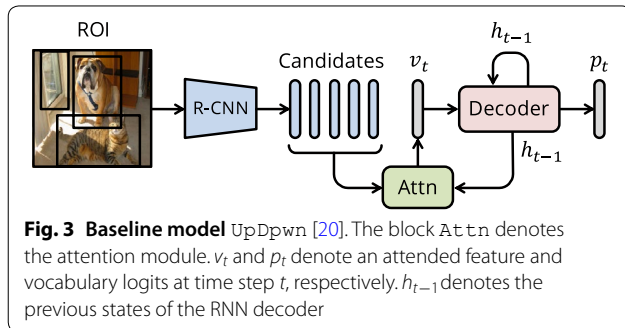
$$V = \{v_1, \dots, v_N\}, v_i \quad (1)$$

We use Faster R-CNN [22] detector, modeled in UpDown [20], to detect objects or salient regions as bounding box assigned to 1,600 object classes and their 400 attribute classes of Visual Genome [23]. The detected raw regions are processed with non-maximum suppression to filter out overlapping, and for each selected region the mean-pooled feature is extracted from the penultimate layer of the object/attribute classifiers as  $v_i$ . Each feature  $v_i$  represents high-level semantic information about the partial scene of the image.

### Salient region clustering

As shown in Fig. 4a, we found even the baseline model can generate reasonable captions by simply bundling the multi-perspective feature sets and being followed by the decoding step (hereinafter called Ensemble), since the attention module can implicitly fuse the correlated ROIs responding to the same top-down signal from the decoder. However, the implicit fusion by the top-down signal may result in biased weights on repetitively occurred objects or may fail to co-occur the ROI features of the identical object due to the subtle difference in the subspace.

Therefore, in this paper, we propose a bottom-up fusion approach of the multi-perspective feature sets. As described in the Sect. “Image encoding”, the image encoder is pre-trained to classify the diverse object classes and attributes so that the encoded ROI features represent well-abstracted semantics. Regarding this property, many studies have demonstrated that CNN-coded features can be diverted for image retrieval tasks due to the fact that semantically similar images are embedded to be close in the learned feature space [24]. Similarly, we assume different views of an identical object are embedded close together in the high-dimensional feature space. Therefore,



our bottom-up approach first clusters the set of multi-perspective features into a number of groups by a standard clustering algorithm such as  $k$ -means [25], as shown in Fig. 4b. When converged, we finally select the centroid vectors as attention candidates. With our assumption, each centroid vector averages a similar set of semantic information across multiple viewpoints. We consistently employ an  $L^2$  distance for inter-sample dissimilarity so that mean-pooled centroids can be adapted to the top-down soft attention. We refer to the reorganized set of features as  $\tilde{V}$ . This clustering approach can easily be extended to a temporal extent so as to smoothen out captions in sequential images. We report the performance improvement by the approach in Sect. “Temporal batch clustering”.

### Caption generation

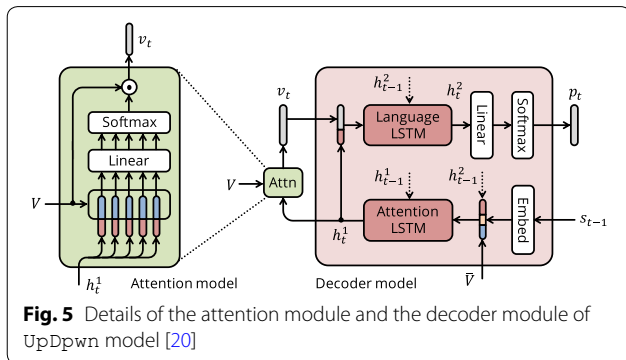
We use the decoder model of UpDown [20] to generate a sequence of words  $S = \{s_1, \dots, s_T\}$  from a set of attention candidate vectors  $\tilde{V}$ . The words  $s_t$  are represented with one-hot vectors where the dimension is equal to the number of vocabulary words  $K$ . The decoder is composed of two stacked long short-term memories (LSTMs) and an attention module, as shown in Fig. 5. At timestep  $t$ , each LSTM updates its hidden state  $h_t$  given previous hidden state  $h_{t-1}$  and an input vector  $x_t$ .

$$h_t = \text{LSTM}(x_t, h_{t-1}) \quad (2)$$

For the input of the first LSTM (“Attention LSTM” in Fig. 5), a previous word  $s_{t-1}$  is first embedded with a learned embedding  $W_e$  and concatenated with a mean-pooled features  $\bar{V}$  and the previous hidden state  $h_{t-1}^2$  of the second LSTM (“Language LSTM” in Fig. 5). We denote the concatenated vector as  $x_t^1$ :

$$x_t^1 = [W_e s_{t-1}, \bar{V}, h_{t-1}^2] \quad (3)$$

Subsequently, the attention module computes a normalized weight  $\alpha_{i,t}$  for each candidate vector  $v_i$  in parallel, given the signal  $h_t^1$  from the first LSTM:



$$e_{i,t} = w_a^\top \tanh(W_v v_i + W_h h_t^1) \quad (4)$$

$$\alpha_{i,t} = \frac{\exp(e_{i,t})}{\sum_{k=1}^{|V|} \exp(e_{k,t})} \quad (5)$$

where  $w_a$ ,  $W_v$ , and  $W_h$  are the learned parameters. The second LSTM then inputs a vector combining the hidden state  $h_t^1$  and the attended feature  $v_t$  is computed as a convex combination of all attention candidates  $v_i$ .

$$v_t = \sum_{i=1}^M \alpha_{i,t} v_i \quad (6)$$

$$x_t^2 = [v_t, h_t^1] \quad (7)$$

Finally, a probability distribution  $p_t$  over the predefined vocabulary words is generated at each time step  $t$ , given by the classifier.

$$p_t = \text{softmax}(W_p h_t^2) \quad (8)$$

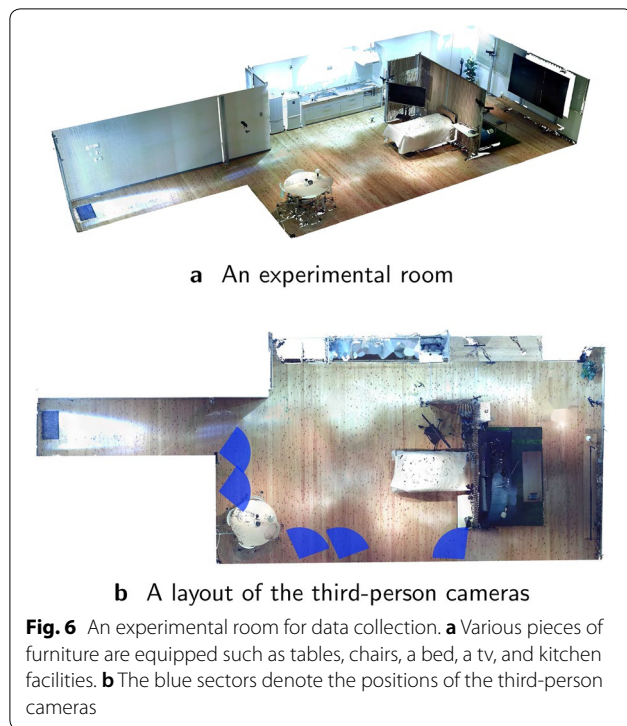
We applied beam-search decoding with a beam-width of 5. Only stopwords such as “a”, “and”, or “the” are allowed to be sampled more than once in decoding.

## Experiments

In this section, we provide an evaluation dataset that we newly constructed, the experiment setups, and the performance.

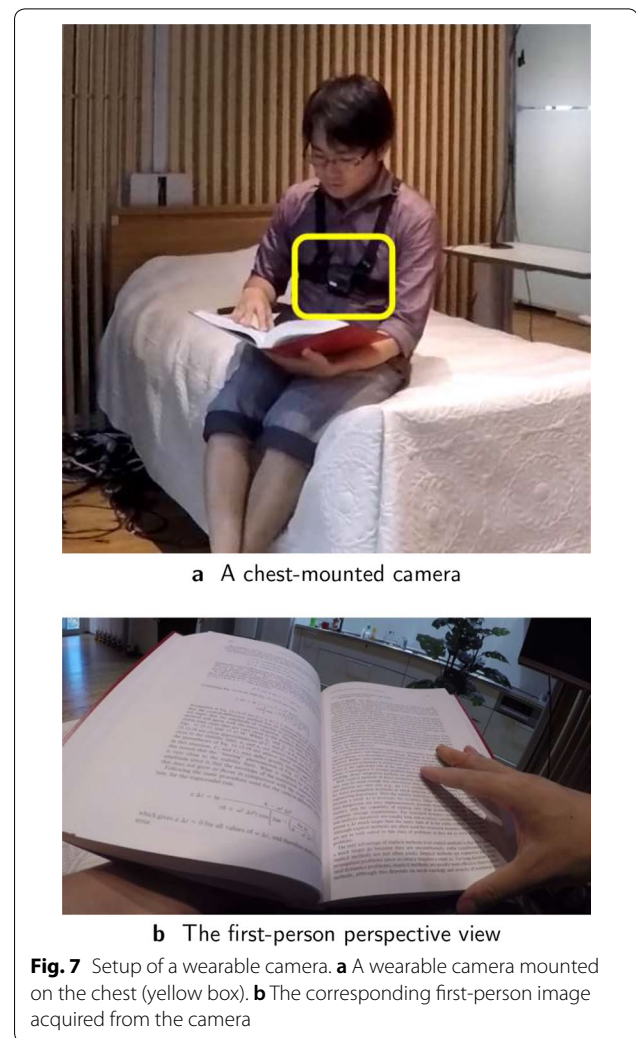
### Dataset

To evaluate the independent effect of each perspective, we collected perspective-aware lifelog videos in our experimental room and corresponding scene descriptions for each video. Fig. 6a shows the appearance of the experimental room to collect the lifelog videos. In the experimental room, various pieces of furniture are equipped such as tables, chairs, a bed, a tv, and kitchen facilities. When recording data, we randomly arranged household items such as bottles, books, cups, and a plant. We employed five participants for the dataset construction and for each recording two out of them were asked to wear action cameras (GoPro HERO). Similar to the related work [8, 9], the action camera was attached to their chests as shown in Fig. 7a, where we found that produced stable results than the head-mounted way. One participant is an actor (a role of a residential person) who was asked to perform everyday activities such as reading books, walking around, washing dishes, and the recorded videos are labeled as the first-person. On the other hand, another participant is an observer (a role of a service robot) who was asked to adaptively monitor the



actor, and the recorded videos are labeled as the second-person. We requested the observer to follow two moving policies: to capture not just the actor in close but also his/her forehead manipulations and to respond to interactions from the first-person actor such as handing objects. Simultaneously, another action camera is fixed on the wall to provide the third-person perspective to capture both participants as shown in Fig. 6b. The position of the third-person camera is randomly selected for each recording. We collected the data for seven rounds with different combinations of 5 participants, while each video is recorded in several seconds. All perspective videos were captured in  $1920 \times 1080$  resolution at 30 fps with  $133.6^\circ$  diagonal angle of view. Subsequently, the videos are downsampled to  $640 \times 360$  with bilinear interpolation. The total number of synchronized video sets is 230 and the total number of frames is 24,587 for each perspective. Examples of the captured images can be seen in Fig. 11. The bottom two examples are sampled from the interactive scenes where the participants are handing over household items.

We built an annotation interface to generate reference captions. As we exclusively focus on modeling an actor's daily living, annotators are asked to describe the actor's activities and interactions viewed through the multi-perspective videos with a single sentence in English. We collect five captions for each video set and the total annotation comprises 1,150 captions. The annotated caption

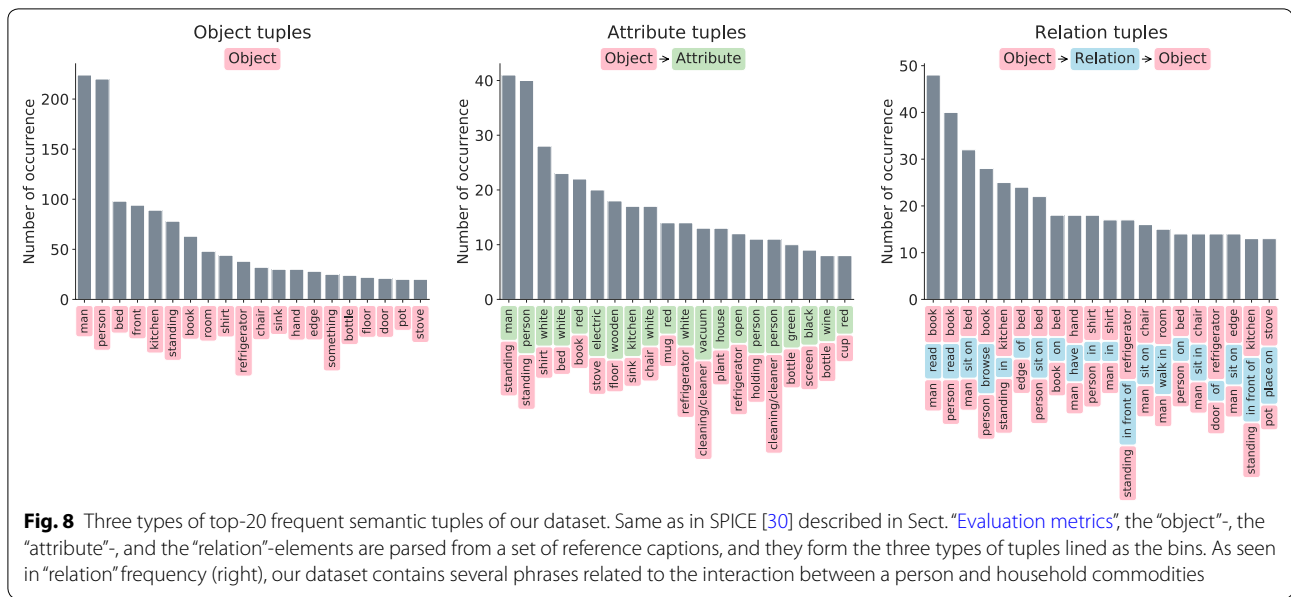


comprises 12.8 words on average. Fig. 8 provides dataset statistics based on semantic tuples frequently occurring in the captions. There are three types of semantic tuples: object tuples, attribute tuples, and relational tuples, which are extracted by parsing the captions. The detailed procedure is described in Sect. "Evaluation metrics". According to the statistics, the annotations cover possible situations and contain several phrases related to the interaction between the actor and the household commodities.

#### Evaluation metrics

For metrics to evaluate the quality of a generated caption against references, we employ the widely used BLEU [26], ROUGE [27], METEOR [28], CIDEr [29], and SPICE [30]. Each metric defines the similarity between a candidate sentence and a set of reference sentences based on word agreement in various granularities. BLEU is a weighted geometric mean of precisions over unigrams (1-grams)





to  $n$ -grams with a penalty for sentence brevity, which was proposed for the machine translation task. ROUGE is an  $n$ -gram based recall proposed for document summarization task. We use a modified version, ROUGE-L, which computes a recall-biased  $F_\beta$ -score based on the longest subsequence matched between a candidate and a set of references. METEOR is a recall-biased  $F_\beta$ -score based on unigrams with a penalty for alignment fragmentation, which was proposed for the machine translation task. The unigram alignment is based on stemming and synonym matching with a large lexical database of English. Besides these metrics, CIDEr and SPICE are explicitly designed for evaluating image captions and are better correlated to human judgments. CIDEr is an average cosine similarity of  $n$ -grams weighted by their Term Frequency-Inverse Document Frequency (TF-IDF). The weights contribute to suppressing the effect of the  $n$ -grams commonly occurring in references such as stop words. We use CIDEr-D that removes stemming and adds a length penalty. SPICE is an  $F_1$ -score based on scene graph tuples. The scene graph is a directed graph parsed from sentences with synonym matching, where an object node is connected by attribute nodes and/or relation nodes. A set of object, attribute, relation tokens forms the tuple, and the similarity score is calculated over the tuples between candidate and reference captions.

### Implementation details

The dataset used for training is the Microsoft COCO dataset [21] that includes 113,287 images for training, 5000 images for validation, and 5000 images for testing,

as defined in Karpathy splits [31]. The baseline model UpDown is trained on pairs of an image and five captions in the training split. The captions do not have any punctuation and are unified in lower case. The vocabulary is pruned by defining any words that have a count less than five as a particular `<unknown>` word. The final vocabulary comprises 10,010 words. We use our PyTorch re-implementation of UpDown model originally written in Caffe [20]. For the Faster R-CNN detector, we choose ResNet-101 [32] as a backbone and perform ROI pooling on `pool5` feature maps so as to encode each region into a 2048- $D$  vector. For each image, we select up to 100 candidate regions according to the predicted scores. Since the number of candidate regions is small, the clustering step had little effect on the whole process time in our experiments. Following the baseline [20], The model is trained with the criterion of minimizing cross-entropy of reference captions and Self-Critical Sequence Training (SCST) [33] that directly optimizes the CIDEr scores of sampled captions. We perform beam search decoding with a beam width of 5 until reaching the end token or the maximum length of 20. We restrict the occurred times of words to one except for stop words.

### Quantitative analysis

In this section, We systematically evaluate the similarity scores between generated and reference captions, the effects of clustering settings, and the performance of extended approach on the temporal extent, using our dataset.

### Perspective ablation

To validate the effects of each perspective, we herein generate captions from different combinations among the three types of perspective images. We hereinafter identify the approaches with a name `<method> - <perspectives>`, e.g., Ensemble-123 denotes the ensemble features from the first-, second-, and third-person perspective images. All generated captions are individually evaluated with reference captions.

In Table 1, we provide the scores of five types of metrics evaluated on our dataset. UpDown [20] is a baseline method which inputs a single-perspective image. Ensemble is a method that bundles attention candidates from two or three perspectives and inputs them to the UpDown decoder. KMeans is our proposed method to construct attention candidates by clustering multi-perspective ROI features into  $k$  groups beforehand and then input the  $k$  centroids to UpDown decoder. We initially set 32 to the number of clusters  $k$ , which is close to the best number of attention candidates reported in UpDown results on COCO dataset [20]. The centroids are initialized with  $k$ -means++ [34] algorithm and iteratively updated until converged. As seen in Table 1, KMeans-123, a proposed method for inputting images from three perspectives shows the best scores in all evaluation indices. Focusing on double input models (middle), the score of the proposed

KMeans is higher than that of Ensemble for any input combination, indicating the effectiveness of bottom-up clustering. We note that the SPICE scores of UpDown-1 and UpDown-2 are close; however, even Ensemble-12 that simply combines the first- and second-person images boosts the performance. It can be considered that each perspective has complementary visual cues to generate actor-related descriptions. This observation can also be seen the most in the case of the third-person perspective although the image itself is confusing to exclusively extract the actor's features. Another important observation is that Ensemble-12 is better than Ensemble-123 while KMeans-123 performs better than KMeans-12 with the adoption of the bottom-up fusion clustering.

Table 2 shows the relative improvement rates of the triple input model against the double input model, in that, the contribution of the ablated perspective. For example, the "First-Person" scores are computed from KMeans-123 and KMeans-23. For precision-based BLEU metrics, the second-person perspective shows the highest rates. Notably, for the other metrics, the first-person perspective shows the highest rates. It can be considered that the second-person images explicitly and exclusively capture the actor's scenes to generate an actor-wise description, but the other important visual cues reside on the other perspective images.

**Table 1** Ablation study of image captioning performance on our dataset

Input perspective	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr-D	SPICE
First	UpDown [20]	51.20	33.47	20.41	11.25	38.85	17.45	21.44	12.19
Second	UpDown [20]	60.86	43.24	31.12	21.19	45.60	19.46	16.94	12.08
Third	UpDown [20]	42.80	26.56	16.17	9.70	31.34	13.73	6.79	6.28
Second Third	Ensemble	59.14	41.97	30.45	21.06	44.09	19.13	15.18	11.40
Second Third	KMeans	62.31	45.34	33.16	22.91	46.22	20.19	17.76	12.21
First Third	Ensemble	59.06	42.78	30.47	20.28	45.16	20.33	27.71	14.37
First Third	KMeans	60.83	44.71	32.03	21.48	46.27	21.16	30.10	15.02
First Second	Ensemble	62.08	45.37	32.82	22.47	47.67	21.68	30.03	15.04
First Second	KMeans	62.43	45.78	32.90	22.19	47.61	21.87	30.76	15.24
First Second Third	Ensemble	63.12	46.37	34.08	23.71	47.92	21.72	29.52	14.99
First Second Third	KMeans	65.09	48.93	36.02	24.78	49.13	22.79	33.41	15.72

Highest values are in italic

**Table 2** The relative improvement rate (%) of triple input vs ablated double input. For example, the "First-Person" scores are computed from KMeans-123 and KMeans-23

	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr-D	SPICE
First-person	KMeans	+4.5	+7.9	+8.6	+8.1	+6.3	+12.9	+88.1	+28.7
Second-person	KMeans	+7.0	+9.4	+12.4	+15.4	+6.2	+7.7	+11.0	+4.7
Third-person	KMeans	+4.2	+6.9	+9.5	+11.7	+3.2	+4.2	+8.6	+3.1

Highest values are in italic

### Detailed results of SPICE

The SPICE  $F$ -score is computed based on semantically parsed sentences and can be subdivided into six meaningful criteria: object, relation, attribute, color, count, and size [20]. Table 3 summarizes the results as the same format as Table 1. Although our triple input model KMeans-123 still outperforms in the object category; in other categories, single- or double-input models are better. The first-person model is remarkably superior to others in the color category. That indicates the joint attention of ROI features across multiple perspectives possibly obscures the detailed visual concepts, meanwhile stably generating captions that includes important concepts such as the actor or other items.

### Clustering setting

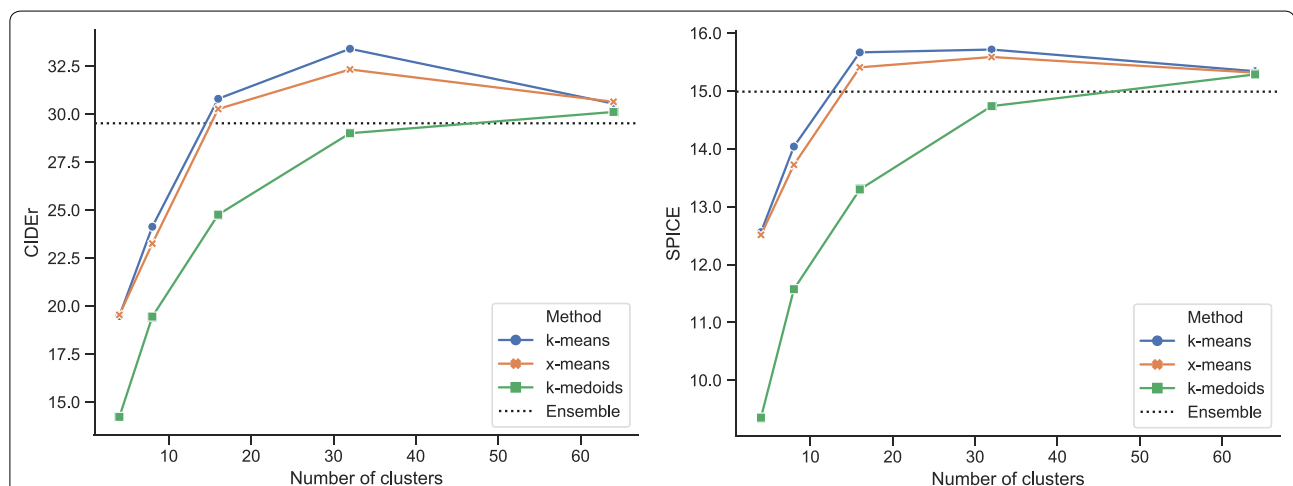
Herein, we report performance comparison in clustering algorithms based on KMeans-123. For clustering

algorithms, we compare basic  $k$ -means [25] that produces  $k$  centroids,  $x$ -means [35] that adaptively subdivides the clusters under the Bayesian information criterion, and  $k$ -medoids [36] that update medoids instead of centroids. All algorithms are initialized with the  $k$ -means++ [34] algorithm. For the number of clusters, we sweep {4, 8, 16, 32, 64} for each algorithm. We note that only  $x$ -means may increase the number in updating, and all algorithms could be equivalent to Ensemble-123 when the number of clusters reaches the number of attention candidates. Fig. 9 shows CIDEr-D and SPICE scores on various methods of clustering algorithms and Ensemble-123 without clustering. For both metrics, the peaks are at  $k = 32$  in  $k$ -means and  $x$ -means algorithms, while the  $k$ -medoids algorithm that “prunes” attention candidates reduces both scores as the number of clusters decreases. Although  $k$ -medoids algorithm is known as robust to noises and outliers, each selected

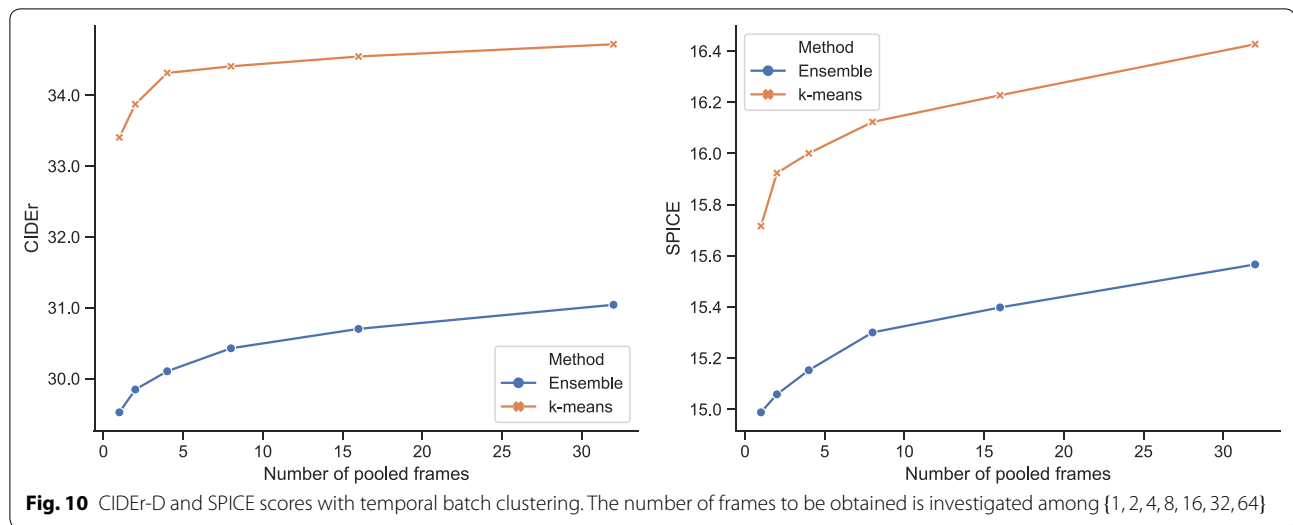
**Table 3** SPICE subcategory scores on our dataset

Input Perspective	Method	SPICE (All)	Object	Relation	Attribute	Color	Count	Size
First	UpDown [20]	12.19	26.36	1.42	3.52	2.38	0.00	0.00
Second	UpDown [20]	12.08	26.66	1.30	1.45	0.02	0.00	0.00
Third	UpDown [20]	6.28	14.62	0.46	0.17	0.04	0.11	0.00
Second Third	Ensemble	11.40	25.42	1.13	0.87	0.00	0.00	0.00
Second Third	KMeans	12.21	27.40	1.08	0.86	0.00	0.00	0.00
First Third	Ensemble	14.37	30.48	2.15	3.30	0.17	0.00	0.00
First Third	KMeans	15.02	32.02	2.13	3.16	0.15	0.00	0.00
First Second	Ensemble	15.04	31.96	1.98	3.63	0.04	0.00	0.00
First Second	KMeans	15.24	32.56	1.90	3.41	0.14	0.00	0.00
First Second Third	Ensemble	14.99	32.02	2.01	3.35	0.02	0.00	0.00
First Second Third	KMeans	15.72	33.74	1.96	3.18	0.05	0.00	0.00

Highest values are in italic



**Fig. 9** CIDEr-D and SPICE scores with different methods of clustering. The number of clusters  $k$  is swept among {4, 8, 16, 32, 64} for each. Both metrics show the best scores at  $k = 32$



medoid is subject to only one salient region while the other algorithms use centroids in which clusters are averaged. We see that indicating roughly grouped centroids

still preserve the visual features and that enables us to encourage implicit joint attention across perspectives and to reflect in the captions.



**Fig. 11** Example captions with the single perspective model and our proposed model. Methods “First”, “Second”, and “Third”: UpDown model with a single image. Method “Ours”: KMeans model with three types of images. All results are generated with beam search decoding



### Temporal batch clustering

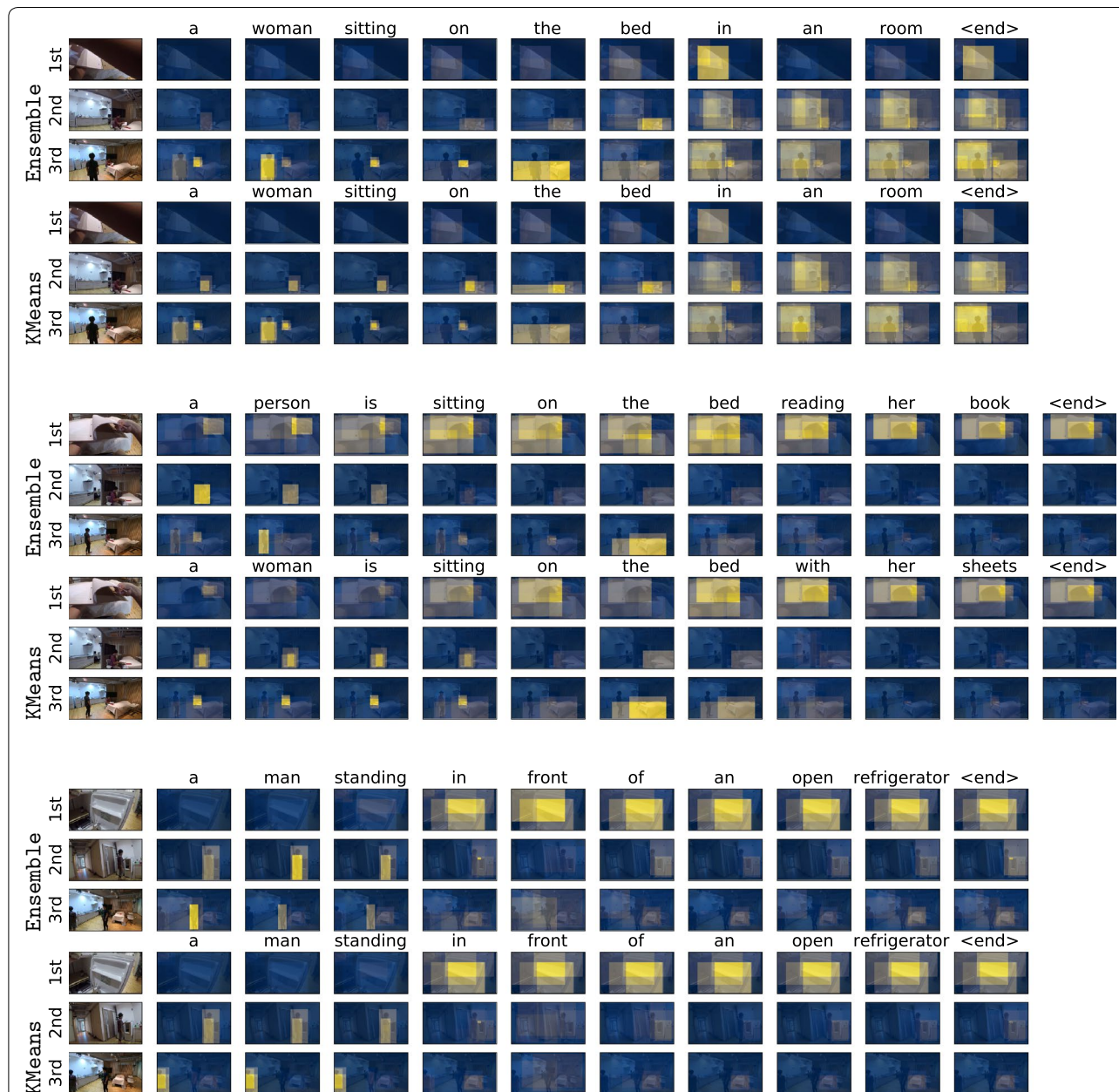
As mentioned in Sect. “Salient region clustering”, the clustering approach can easily be extended to a temporal extent so as to smoothen out captions in sequential images. Herein, we report the performance of KMeans-123 and Ensemble-123 where the attention candidates are aggregated across the consecutive frames and not just only perspectives. For the number of pooled frames, we sweep {1, 2, 4, 8, 16, 32} for both approach. As

seen in Fig. 10, both approaches boost the performance as the number of frames increase.

### Qualitative analysis

#### Generated captions

In Fig. 11 we show some caption examples of the first-person (UpDown-1), second-person (UpDown-2), third-person models (UpDown-3), and our fourth-person model (KMeans-123). The bottom two examples show



**Fig. 12** Top-down attention maps of KMeans and Ensemble. The yellow rectangular regions have a large attention weight and the group of regions in column conditions the decoder in generating each word. It can be observed that our proposed method KMeans discriminates instances against Ensemble. Best viewed in color

interactive scenes where the participants are handing over household items. The first-person perspective captions tend to briefly describe the relationship of the actor and objects manipulated by hands, whereas there is no explicit phrase about context such as the type of place. In contrast, the second-person perspective captions succeed at describing where the actor is and his/her postures, which are invisible concepts from a first-person perspective. However, in some cases, the types of activities are not clear due to their visual granularity. The third-person perspective captions are more ambiguous in terms of the participants' situation, while novel objects not visible in other perspectives are described. Finally, our proposed approach generates more detailed captions about the actor and the context. For instance, on the top left figure, the caption by our method includes the actor's posture, location, and detailed activity, which could only be described partially through single perspective cases. Moreover, in interactive cases on the bottom, we can see that our proposed method improves the third-person captions with the additional phrase about the manipulated objects derived from the first-person and/or the second-person image sources. Qualitatively, we see that the critical visual concepts are potentially in the first- and second-person perspectives, while the third-person perspective contributes to describe the interactive scenes. Although the verbal expression is slightly different for each perspective, we can see our method successfully produced reasonable description integrating three types of perspectives semantically. It can be considered that our clustering on ROI feature space effectively works to summarize the multi-perspective images.

### Visualizing ROI attention

To verify how salient regions are attended in decoding each word, we here visualize the attention weights  $\alpha$  in Eq. 6 associated with ROIs. We colorize all detected salient regions according to the attention weights. The weights vary in each decoding step and show us which regions are contributed to predicting each word. In Fig. 12, we provide three instances each of which has two types of results from Ensemble-123 and KMeans-123. We can see remarkable differences in predicting human-related words. Ensemble-123 model incorrectly focuses on different people or either of which, while our KMeans-123 model successfully focuses on the same person in the second- and third-person images. That indicates our proposed bottom-up approach is effective to improve instance-correspondence in the top-down weighting.

## Conclusion

In this paper, we proposed a novel approach to automatically generate lifelog captions using multi-perspective images in human–robot symbiotic scenarios, these involve, the first-person image from a wearable camera, the second-person image from a camera mounted on a robot, and the third-person image from an embedded camera in the intelligent space. To generate captions from the triplet images, we introduced a bottom-up fusion step that clusters salient region features across the perspectives, and we integrated it with the attention-based image captioning architecture. Next, in a living environment with furniture and daily necessities, we collected the fourth-person lifelog images and multiple reference captions assuming human–robot symbiosis scenarios. Finally, we demonstrated the effectiveness of our proposed approach through perspective-aware ablation studies. The results showed that simply increasing and decreasing candidates do not contribute to the caption scores in our multi-perspective setting, as demonstrated in our experiments of perspective ablation and clustering modification. A limitation is in that our mechanism still does not have a function to explicitly specify the human regions to be described. Future work includes the development of a compositional captioning method to exclusively control visual attention to avoid confusing a human with a robot. Moreover, we will scale-up the types of environments to be recorded and evaluate the approach in terms of lifelogging searchability.

### Acknowledgements

Not applicable.

### Authors' contributions

KN carried out the main part of this study and drafted the manuscript. YI and RK contributed to the fundamental concept of this study and revised the manuscript. All authors read and approved the final manuscript.

### Funding

This work was supported by Grant-in-Aid for JSPS Fellows Grant Number 19J12159 and JST CREST Grant Number JPMJCR17A5.

### Availability of data and materials

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup> Graduate School of Information Science and Electrical Engineering, Kyushu University, 744 Motooka Nishi-ku, Fukuoka-shi, Fukuoka 819-0395, Japan. <sup>2</sup> Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr, Pasadena, CA 91109, USA. <sup>3</sup> Faculty of Information Science and Electrical Engineering, Kyushu University, 744 Motooka Nishi-Ku, Fukuoka-shi, Fukuoka 819-0395, Japan.

Received: 27 January 2020 Accepted: 18 September 2020  
Published online: 24 September 2020

## References

- Hodges S, Williams L, Berry E, Izadi S, Srinivasan J, Butler A, Smyth G, Kapur N, Wood K (2006) SenseCam: A retrospective memory aid. *Proceedings of the International Conference of Ubiquitous Computing (UbiComp)*, pp. 177–193
- Bolaños M, Dimiccoli M, Radeva P (2016) Toward storytelling from visual lifelogging: an overview. *IEEE Trans Hum Mach Syst* 47(1):77–90
- Xiong B, Grauman K (2014) Detecting snap points in egocentric video with a web photo prior. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 282–298
- Poleg Y, Arora C, Peleg S (2014) Temporal segmentation of egocentric videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2537–2544
- Singh S, Arora C, Jawahar C (2016) First person action recognition using deep learned descriptors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2620–2628
- Bettadapura V, Essa I, Pantofaru C (2015) Egocentric field-of-view localization using first-person point-of-view devices. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 626–633
- Pirsiavash H, Ramanan D (2012) Detecting activities of daily living in first-person camera views. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2847–2854
- Fan C, Zhang Z, Crandall DJ (2018) Deepdiary: lifelogging image captioning and summarization. *J Visual Commun Image Representation* 55:40–55. <https://doi.org/10.1016/j.jvcir.2018.05.008>
- Bolaños M, Peris Á, Casacuberta F, Soler S, Radeva P (2018) Egocentric video description based on temporally-linked sequences. *J Visual Commun Image Representation* 50:205–216
- Lee J-H, Hashimoto H (2002) Intelligent space-concept and contents. *Adv Robotics* 16(3):265–280
- Kurazume R, Pyo Y, Nakashima K, Kawamura A, Tsuji T (2017) Feasibility study of iort platform “big sensor box”. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3664–3671
- Nakashima K, Iwashita Y, Yoonseok P, Takamine A, Kurazume R (2015) Fourth-person sensing for a service robot. *Proceedings of the IEEE Conference on Sensors*, pp. 1–4
- Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems (NIPS)*, pp. 568–576
- Clinch S, Metzger P, Davies N (2014) Lifelogging for ‘observer’ view memories: an infrastructure approach. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pp. 1397–1404
- Sigurdsson GA, Gupta A, Schmid C, Farhadi A, Alahari K (2018) Actor and observer: joint modeling of first and third-person videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7396–7404
- Xu M, Fan C, Wang Y, Ryoo MS, Crandall DJ (2018) Joint person segmentation and identification in synchronized first-and third-person videos. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 637–652
- Fan C, Lee J, Xu M, Kumar Singh K, Jae Lee Y, Crandall DJ, Ryoo MS (2017) Identifying first-person camera wearers in third-person videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5125–5133
- Vinyals O, Toshev A, Bengio S, Erhan D (2017) Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans Pattern Anal Mach Intell (TPAMI)* 39(4):652–663. <https://doi.org/10.1109/TPAMI.2016.2587640>
- Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2048–2057
- Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6077–6086
- Chen X, Fang H, Lin T-Y, Vedantam R, Gupta S, Dollár P, Zitnick CL (2015) Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*
- Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NIPS)*, pp. 91–99
- Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma DA et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vis (IJCV)* 123(1):32–73
- Zheng L, Yang Y, Tian Q (2017) SIFT meets CNN: a decade survey of instance retrieval. *IEEE Trans Pattern Anal MachIntelligence (TPAMI)* 40(5):1224–1244
- Lloyd S (1982) Least squares quantization in PCM. *IEEE Trans Inform Theory* 28(2):129–137
- Papineni K, Roukos S, Ward T, Zhu W-J (2002) BLEU: a method for automatic evaluation of machine translation. *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, pp. 311–318
- Lin C-Y (2004) ROUGE: A package for automatic evaluation of summaries. *Proceedings of the ACL Workshop on Text Summarization Branches Out*, pp. 74–81
- Banerjee S, Lavie A (2005) METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization*, pp. 65–72
- Vedantam R, Lawrence Zitnick C, Parikh D (2015) CIDEr: Consensus-based image description evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
- Anderson P, Fernando B, Johnson M, Gould S (2016) SPICE: Semantic propositional image caption evaluation. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 382–398
- Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778
- Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V (2017) Self-critical sequence training for image captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7008–7024
- Arthur D, Vassilvitskii S (2007) k-means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035
- Pelleg D, Moore A (2000) X-means: Extending k-means with efficient estimation of the number of clusters. *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 727–734
- Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. *Math Intelligencer* 27(2):83–85