Improving Point of View Scene Recognition by Considering Textual Data

Volkmar Frinken, Yutaro Iwakiri, Ryosuke Ishida, Kensho Fujisaki, Seiichi Uchida
Faculty of Information Science and Electrical Engineering
Kyushu University, Japan
{vfrinken, uchida}@ait.kyushu-u.ac.jp, {iwakiri, ishida, fujisaki}@human.ait.kyushu-u.ac.jp

***Abstract*—At the current rate of technological advancement and social acceptance thereof, it will not be long before wearable devices will be common that constantly record the field of view of the user. We introduce a new database of image sequences, taken with a first person view camera, of realistic, everyday scenes. As a distinguishing feature, we manually transcribed the scene text of each image. This way, sophisticated OCR algorithms can be simulated that can help in the recognition of the location and the activity. To test this hypothesis, we performed a set of experiments using visual features, textual features, and a combination of both. We demonstrate that, although not very powerful when considered alone, the textual information improves the overall recognition rates.**

I. Introduction

Wearable camera devices are widely being discussed due to their social acceptance, but also because of their potential to facilitate peoples' lives. Such cameras record a video or image sequence from the user's perspective, and are frequently called *First Person View* (FPV), *Point of View* (PoV), or *egocentric* (EC) cameras.

Sophisticated usage of these cameras rely heavily on pattern recognition algorithm to extract useful information from the sequence of images. Having a camera recording current scenes, e.g, makes to possible to automatically provide useful information, record key events for future reference, or share experiences with other people. Hence, corresponding algorithms need to be able to recognize what the user is doing and where he or she is.

The visual representation alone, however, might not be enough for robust recognition of the recorded scenes. A natural extension to this is to use the texts which occur in the scenes. Unfortunately, scene text detection and recognition is a hard problems, with intense ongoing research [7], [14].

In order to promote this research area and to facilitate related investigations, we present the PoVIST DB, a *Point of View Image and Scene Text Database* to investigate how textual data can be useful for the recognition without having to wait for the creation of robust and well-performing OCR systems. The PoVIST DB database provides a manual transcription along with each image. With these data, we propose to perform two recognition tasks, viz. the location of the user wearing the camera and the action performed by the user.

Several databases exist that cover related research areas, yet no database has been published yet, to the knowledge of the authors, that contains an egocentric camera view of with a textual ground truth of all the text images. In [11], [13] databases are presented for activity recognition based on videos. or shot from a wearable device. None of the above mentioned databases are recorded with a PoV camera.

As far as databases are concerned that contain videos or image sequences shot from an egocentric camera, in [8] a database of different sport activities is proposed. In [3] detailed actions of the protagonist, like preparing food, are to be recognized and in [10] various household activities should be classified. In a recent database, proposed in [12], humans are asked to interact with a statue wearing a camera.

Thus, the first contribution of this paper is the PoVIST DB, which, in contrast, was recorded in different settings (outside, inside, city, rural areas, etc.) and provide a manually labeled transcription of all text that occurs in each frame. This way, an OCR step can be simulated to develop and compare algorithms that make use of textual as well as visual features.

The second contribution of this paper is a set of experiments to show the importance of the scene text and to provide baseline results for future references. First, we performed action and location classification using common visual features with a state-of-the-art neural network-based sequential classifier. Then we extracted textual features using latent semantic analysis and repeated the process. It turns out that the textual information substantially improves the recognition performance.

The rest of this paper is structured as follows. In Section II, the motivation for creating a new data set are explained. The details and statistics of PoVIST DB along with two tasks associated with the database are presented in Section III. The recognition systems are explained in Section IV. Baseline results are given in Section V and conclusions are drawn in Section VI.

II. Motivation

Wearable PoV cameras are relatively new data source which has the potential to be closely integrated in our daily lives. One can imagine a camera attached to pieces of clothing or accessory that take videos or pictures roughly towards the directions towards which the user (the protagonist) directs his or her gaze, but without any further information about the center of attention. This poses obviously a severe set of challenges to the pattern recognition algorithm of a processing tool-chain. For example, they need to recognize the scene or

identify the region of interest, before any further steps can be taken.

With such a setup in mind, we created a database that fulfills the following requirements. First, the data should be recorded in a natural manner, i.e. we did not set up any scenes. Every picture was recorded without any specific instructions given. Second, the data should be realistic, i.e. not every image might contain useful information and the regions of interest are embedded in substantial background noise. Third, actions performed and locations visited can sometimes be ambiguous. Forth, *interesting* actions where a wearable device could be useful, like shopping or visiting a museum, should be mixed with *uninteresting* actions, i.e. one from which a system can not directly infer information. Fifth, and finally, the database should not just provide a final ground truth for training and testing, but also the intermediate information of a manual transcription.

With a database containing natural scenes we want to provide real data that reflects what a user of an egocentric camera would experience when wearing it in the public. Therefore, we did not prepare any scenes but defined only vague tasks, such as "shopping in a downtown area". This way, we want to make sure that the data is natural and therefore representative enough to develop algorithms that can be used in real word tasks.

A similar argument is to be made for the second point. Realistic predictions about the applicability can only be made when algorithms are robust enough to deal with a substantial amount of noise, up to the point that some images might not contain useful information at all and context from previous images has to be used.

In order to not give any bias towards the database, we defined the actions performed and the locations of each image after the data has been recorded. Hence, some image are taken of scenes which are ambiguous and can not clearly be assigned to an action or location class.

Also, images exist in which no clear action is performed at all. Yet, this reflects a real word scenario better than if, for instance, the protagonist goes straight from one, clear action to the next one or turns the device on only during the time he or she is performing a clear action. Instead, a useful algorithm should automatically detect whether useful information is contained in the recorded image or not.

The most prominent features that sets this database apart from other databases, however, is that of the provided transcription. Algorithms for scene text detection and recognition exist that work to a certain degree [7], [14]. In can be assumed that in the future, the performance might increase to be close to that of humans. At this point, visual features extracted from the images can be complemented with textual features. Accessing not only the visual features but also the semantic content of the texts might help the activity and location recognition. Thus, the PoVIST DB provides both sources. The experimental evaluation presented in Section V demonstrates that the textual data provides valuable information that helps to recognize the scenes recoded by the camera.

TABLE I
STATISTICS ABOUT THE POVIST DB.

Recording Session	Number of Images	Sub-sequences with constant locations and activity	Average length of constant sub-sequence
University Campus	1 621	264	6.14 frames
Rural Village	2 959	210	14.09 frames
Downtown Shopping	1 935	177	10.93 frames

In addition, the sequential nature of the recorded images also creates dependencies between consecutive images. In fact, some activities can not be recognized confidently with a single image but only when observing an image in context with preceding ones. For an interactive system, it might be important to differentiate between walking through a store, looking briefly at a shelf, or intensely studying the different products.

III. THE DATABASE

We have chose to take images in regular intervals from a camera without further specifying any restrictions, such as the focus, region of interest, etc. while performing some actions. The database contains images from three recording sessions with the protagonist performing different tasks with an overall duration of over 18 hours. The recording was done with a camera that was hanging around a persons neck which automatically took pictures every 10 seconds. In the first session (University Campus), the protagonist walks around the local university campus. In the second session (Rural Village), the protagonist drives to a small town, goes to a restaurant, visits a museum, etc. In the third session (Downtown shopping), the protagonist goes shopping in a variety of different shops in the shopping district of a large city. All image sequences are taken in or around Fukuoka, Japan.

A. Ground Truth

For each image, three different types of manually annotated ground truth is provided. These are the activity, the location, and the transcription of the text in the scene. The taxonomy of the possible classes of the activity and the location are arranged in a tree, to allow different levels of detail to provide a flexible classification system. For some applications it might, e.g., be enough to differentiate between outside and inside, while for others the type of shop visited might be important as well. The list of available activities is given in Table III. On the coarse scale, the activities are split into *moving*, *stopping*, and *other*. On a finer scale, moving is sub-divided into *driving* and *walking*. Similarly, the different locations are given in Table II. This class tree is larger and provides a rough classification into *inside* and *outside* on the top level, but get more complex for higher levels of detail.

The third type of ground truth is the textual data seen in the image, provided in a utf-8 text file, as can be seen in the Fig. 1. Since the recording was done in Japan, they contain a large set of Japanese characters. To ease the handling, we

TABLE II
THE HIERARCHY OF POSSIBLE LOCATIONS.

Level 0	Level 1	Level 2	Level 3
Outside	Countryside	Point of Interest	Museum / Exhibition Map / Information Monument Scenic view
		Vendor	Vending Machine Food stand
		Other (Street, etc.)	
	City	Point of Interest	Museum / Exhibition Map / Information Monument
		Vendor	Vending Machine Food stand
		Other (Street, etc.)	
Inside	Shop	Electronic store	Phones Computer
		Supermarket Convenient store CD shop Musical Instrument shop Book shop Stationary	
	Point of Interest	Museum / Exhibition Map / Information	
	University	Floor / Hall Class Room Library	
	Restaurant		
	Mall		
	Office		
	Other		

TABLE III
THE HIERARCHY OF POSSIBLE ACTIVITIES.

Level 0	Level 1
Moving	Driving Walking
Stopping	
Other	

(a) Sample images of the database

Moving
→ Walking

(b) Activity Ground Truth

Outside
→ City
→ Other

(c) Location Ground Truth

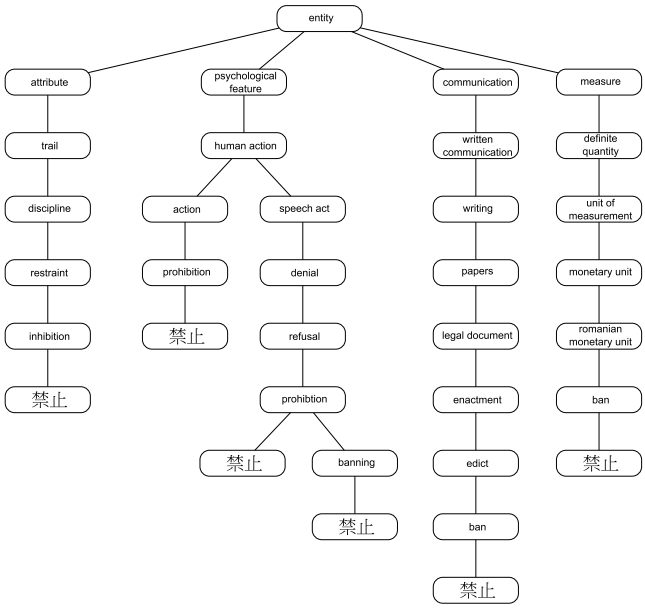
Fig. 1. Four sample images (a) and the provided activity ground truth (b) and the location ground truth (c) for the top left image.

？イ捨て禁止
混み合う街中では
〝おしチャリ〟
おっと危ない！
人混みの中の
そのスピード
P
区画線内の
自転車に限る
ここまで

(a) Text Ground Truth

禁止, 街, 中, お, チャリ,
おっと, 人, 中, スピード,
区画, 線, 内, 自転車, ここ

(b) The list of words



(c) WordNet tree of the first word in the list

Fig. 2. The provided textual ground truth (a), the list of words after the word stemmer (b) and the WordNet tree of the first word in the list (c).

provide therefore not only the original ground truth, but also the list of individual words, extracted by an automatic word stemmer, as well as the translation into English via links into the corresponding branches of the English WordNet [5], [9].

A quantitative summary of the database is given in Table I, which shows for each recording session the number of images taken, the numbers of sub-sequences with a constant location and activity ground truth and the average length of such a sub-sequence. The images were taken with a regular interval of 10 seconds. The first session produced 1 621 images, the second session produced 2 959 images and the third session produced 1 935 images.

In Fig. 1, a sample image from the downtown shopping session is shown. The activity ground truth is "Moving → Walking" and the location is "Outside → City → Other (Street, etc.)" as given in Fig. 1(b-c). The text transcribed by manual annotation is shown in Fig. 2(a). Note, that some of the characters can not be read in the image. These characters are transcribed with the symbol '?'. Upon careful observation, characters can also sometimes be found in the images that escaped the annotators' eyes and do not occur in the tran-

scribed text at all. After applying a word stemmer that also removes stop-words, a list of symbols remains (Fig. 2(b)). For each entry in the list, all possible translations can be arranged in a tree, according to WordNet. How such a tree looks like is shown in Fig. 2(c) where the first word of that list is given. By providing the stemmed word list as well as the corresponding entries in the English WordNet tree, the transcription In fact can just be treated as abstract symbols. This renders the database useful without the requirement of knowing Japanese.

IV. Features and Recognition Systems

To recognize the location and the activity we used Long Short-Term Memory Neural Networks with features extracted from the visual representation only, the textual representation, and both. For each setup, we decided on a few parameters to be optimized on a validation set.

A. Visual Data

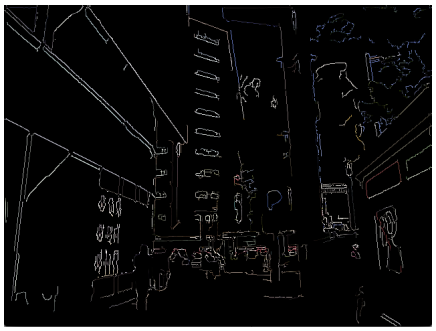
Visually, an image is represented by gradients along edges and color information. For the gradient information, we used an Canny edge detector [1] to cancel out noise and focus on the edges. Afterwards, a dense grid of histogram of oriented gradients (HOG) is used to extract 9 gradients at each cell. After resizing the image to 640×480 pixels, we used a HOG block size of 80×80 pixels, a cell size of 40×40 pixels and a block stride of 40 pixels in both dimensions. This results in a 5940-dimensional vector. This vector is augmented by the average RGB color information of each cell, resulting in a vector with a total dimensionally of 6 516. Obviously, using such a vector directly is infeasible. Hence we reduced the dimensionality via PCA, considering all the images of the training set. The resulting size of the output dimensionality was chosen to be an external parameter to be validated. We tested $k_{visual} \in \{10, 15, 25, 50\}$.

B. Textual Data

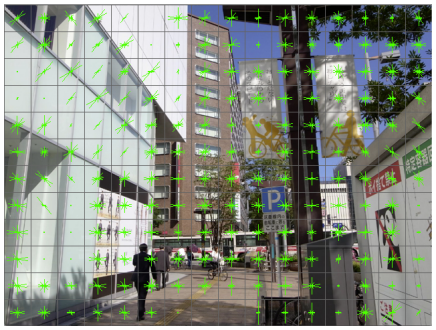
The textual data is represented as a real vector by means of Latent Semantic Analysis (LSA) [2]. First, a term-document matrix $M = (m_{ij})$ is created, where m_{ij} is the term frequency-inverse document frequency (tf-idf) normalized value that indicates how often term t_i occurs in image I_j. Then, M is decomposed via singular value decomposition $M = U\Sigma V^T$, where U, and V and unitary matrices and Σ is a diagonal matrix containing the singular values of D in decreasing order. It is well known that the matrix $M_k = U_k \Sigma_k V_k^T$, where only the top k rows and columns are non-zero, is the best rank k approximation to M, wrt. the Frobenius norm. In LSA, this is used to represent a document as a vector of d normalized term frequencies which is then transformed via $d_k = U_k^T \Sigma_k^{-1} d$ into its k dimensional approximation. Note, that this representation does not take advantage of the WordNet representation, but only of the list of symbols returned after the word stemmer.

The advantages of LSA are the simplicity and of the model which results in a comprehensible, low-dimensional representation of the relative term-frequencies that occur in a document. The dimensionality k is also an external parameter that was optimized on the validation set, using $k_{\text{text}} \in \{10, 15, 25, 50\}$.

(a) Input Image



(b) Canny Edge Detector



(c) Dense HOG Values

Fig. 3. The HOG representation of each images. For the visual representation, the output vector is extended by the average color of each cell and mapped to a lower dimensional vector via PCA.

C. LSTM NN Recognition Approach

To exploit the sequential nature of the images, a recently proposed neural network-based recognizer was used that has been shown to perform extremely well in a large variety of problems, called long short-term memory (LSTM) neural networks. These are recurrent neural networks, endowed with a memory cell in the hidden layer which enables them successfully learn long-term context dependencies [6]. The memory cell is made up from nodes that realize addition, multiplication, and conventional non-linear activation functions. Since

TABLE IV
THE PROPOSED SETS OF THE PoVIST DB.

Set	Blocks	Sub-sequences	Frames
Training Set	50	479	4 269
Validation Set	12	84	1 024
Testing Set	14	88	1 222

all functions are differentiable, LSTM neural networks can be conveniently using back-propagation, respectively back-propagation through time (BTT) for sequential data.

For training and testing, a sequence of vectors is presented to the LSTM network. The input layer has therefore as many nodes as the dimensionality of the chosen representation. The number and sizes of the hidden LSTM layers is an external parameter. The output layer has one node for each possible target class. The output activation levels are finally normalized via *softmax* so that they can be seen as posterior probabilities. The network is trained the target to minimize the cross entropy error [4]. The size h of the hidden layer was evaluated among $h \in \{10, 15, 25, 50\}$. For all experiments, we created 10 randomly initialized LSTM neural networks and report the results of the best network, according to the validation set.

V. EXPERIMENTAL EVALUATION

The recognition performance was evaluated for six different setups, namely using visual data, textual data, or both, for activity and location recognition. Textual and visual data are both represented as n_{text}, resp. n_{visual} dimensional vectors. Since the sequential information is a crucial factor of the database, we use a recognizer that has been designed for sequential data.

A. Database Subsets

The database into a training set, a validation set, and a testing set as follows. Each recording session produced one long sequence of images recoded over the entire time. After labeling each image with an activity and a location, we identified 651 sub-sequences of constant activity and location according to the ground truth. Then, we grouped consecutive sub-sequences into larger non-overlapping blocks with a minimal length of 50 frames. An overview of this approach if shown in Fig. 4. The figure shows the activity and the location for a several consecutive frames and how they are assigned into blocks. Note, that this assignment is purely for the sake of splitting the dataset randomly into training data, validation data, and test data. The final task is frame classification and the recognition system is agnostic to the method how the blocks are created. The final recognition works accordingly on any image sequence.

Finally, all blocks from all sessions are randomly assigned to the training set, the validation set, or the testing set. As can be seen from the Table IV, 50 out of the 75 blocks are used for training, 12 for validation and 14 for testing.

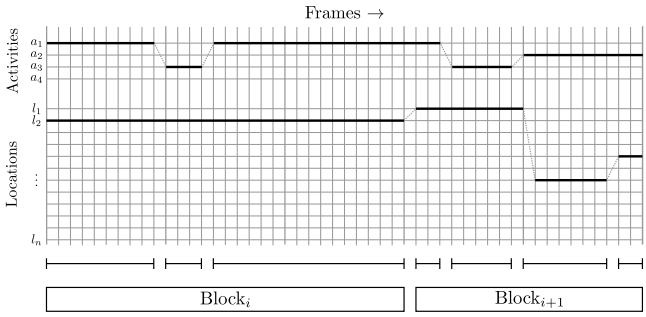


Fig. 4. For training and testing, the images are separated into constant activity and location sub-sequences and then concatenated into larger blocks.

TABLE V
THE RECOGNITION ACCURACIES (%) FOR THE ACTIVITY RECOGNITION.

Set	Features		
	Textual	Visual	Combined
Validation Set	56.93	66.11	66.89
Testing Set	32.56	72.67	79.54

B. Results

In Table V the performance for the activity recognition is given. One can see, that the visual features are far more suitable for the activity recognition than the textual features. However, the best recognition rate can be reported using a combination of visual and textual features.

For the location recognition, a similar situation can be observed as shown in Table VI. The textual features alone are not sufficient to achieve a good recognition rate. In comparison, a better performance is achieved using only the visual features. But again, a combination of both types of features outperforms both systems.

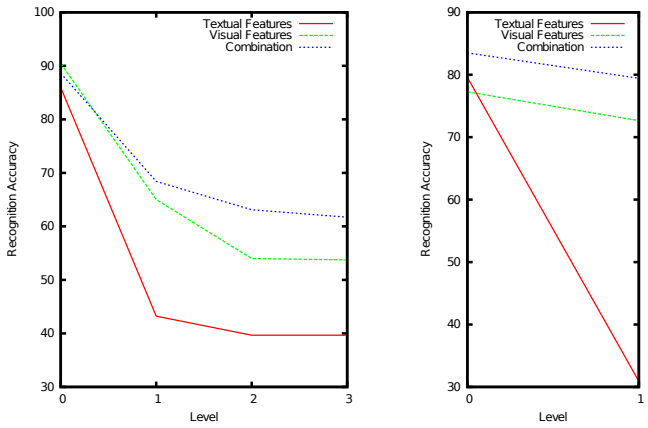
In Fig. 6, samples of corrected recognition due to the influence of the textual features are shown. In Fig. 6a the location recognition without textual features was “Mall” and in Fig. 6b “Other (Street, etc.)”.

The recognition performances regarding the different detail levels are shown in Fig. 5. Especially for the location recognition, it can be seen that the textual data is especially useful for a discrimination on a finer level of detail, while on a coarser scale pure visual features seem to be sufficient.

As far as the validated parameters are concerned, for the activity recognition, the best performance using the textual features is achieved with a 50-dimensional representation and 50 LSTM nodes in the hidden layer of the neural network. Using the visual representation, also, a 50-dimensional vector,

TABLE VI
THE RECOGNITION ACCURACIES (%) FOR THE LOCATION RECOGNITION.

Set	Features		
	Textual	Visual	Combined
Validation Set	51.75	59.96	64.64
Testing Set	39.52	53.93	60.88



(a) Location Recognition

(b) Activity Recognition

Fig. 5. The performance of the best systems as a function of the detail level of the target classes.

(a) Book shop

(b) Map / Information

Fig. 6. Samples of classification results that improved by adding textual features.

but only 10 LSTM nodes lead to be the highest recognition rate. For the combination of both representations the best setup was gain with 10 LSTM nodes and a 50-dimensional input vector, evenly split into 25 textual and 25 visual features. The good performance with just 10 LSTM nodes may reflect the relative easy task of classifying just four different activities, while the large size of the input vectors indicate the difficulty of finding a good representation.

Regarding the location recognition, 35 textual features and 25 LSTM nodes leads to the best recognition rate. The best system using only the visual features used 50 input dimensions and 25 LSTM nodes, whereas the combination of both features gave the highest accuracy when using 35 textual dimensions, 50 visual dimensions and 50 LSTM nodes. Opposed to activity recognition, a larger number of LSTM nodes supports might be needed for the more complex task of location recognition.

VI. Conclusion

Current trends of technology go towards a greater integration of technology in general and cameras in particular in our lives. This offers new possibilities and challenges for pattern recognition algorithms to understand scenes and intentions of the person wearing a camera.

To foster research in this areas, we introduce a novel database for activity and location recognition of a person wearing an egocentric camera that takes pictures in regular intervals. As opposed to existing egocentric camera databases, we provide for every frame not only the activity and location ground truth, but also a manual transcription of the text found. While not always perfect, it can be used black box which provides sophisticated OCR results, which can be expected in a few years, given the current rate of progress in scene text recognition. To facilitate the text processing, word stemming and WordNet entries are provided for every word in every frame as well.

We demonstrate in a set of experiments the advantages of using the textual data for classification. Both tasks, activity and location recognition, profit substantially from considering both input sources. Since future OCR systems are likely to provide comparable transcription results, this database can be used to develop possible algorithms for real world tasks.

The database will be made freely available though the institute website. We are certain that it will be helpful and we hope for a widespread use in wearable camera research.

Acknowledgment

This work is supported in part by the CREST project from Japan Society for the Promotion of Science (JSPS).

References

[1] John Canny. A Computational Approach to Edge Detection. *TPAMI*, pages 679–698, 1986.

[2] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 4(6):391–407, 1990.

[3] Alireza Fathi, Ali Farhadi, and James M. Rehg. Understanding Egocentric Activities. In *ICCV*, 2011.

[4] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer Heidelberg New York Dordrecht London, 2012.

[5] Gao Huini Eshley Hazel Mok Shu Wen and Francis Bond. Using Wordnet to Predict Numeral Classifiers in Chinese and Japanese. In *Global WordNet Association*, 2006.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[7] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazán, and Lluis Pere de las Heras. ICDAR 2013 Robust Reading Competition. In *Int'l Conf. on Document Analysis and Recognition*, pages 1115–1124, 2013.

[8] Kris Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast Unsupervised Ego-Action Learning for First-Person Sports Videos. In *CVPR*, 2008.

[9] George A. Miller. A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[10] Hamed Pirsiavash and Deva Ramanan. Detecting Activities of Daily Living in First-Person Camera Views. In *CVPR*, 2012.

[11] Michael S. Ryoo and Jake K. Aggarwal. Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. In *ICCV*, 2009.

[12] Michael S. Ryoo and Larry Matthies. First-Person Activity Recognition: What Are They Doing to Me? In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, June 2013.

[13] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing Human Actions: A Local SVM Approach. In *17th Int'l Conf. Pattern Recognition*, pages 32–36, 2004.

[14] Asif Shahab, Faisal Shafait, and Andreas Dengel. ICDAR 2011 Robust Reading Competition - Challenge 2: Reading Text in Scene Images. In *Int'l Conf. on Document Analysis and Recognition*, pages 1491–1496, 2011.