Visual Search Target Inference using Bag of Deep Visual Words

Sven Stauden, Michael Barz, and Daniel Sonntag

German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
{sven.stauden,michael.barz,daniel.sonntag}@dfki.de

Abstract. Visual Search target inference subsumes methods for predicting the target object through eye tracking. A person intents to find an object in a visual scene which we predict based on the fixation behavior. Knowing about the search target can improve intelligent user interaction. In this work, we implement a new feature encoding, the *Bag of Deep Visual Words*, for search target inference using a pre-trained convolutional neural network (CNN). Our work is based on a recent approach from the literature that uses *Bag of Visual Words*, common in computer vision applications. We evaluate our method using a gold standard dataset. The results show that our new feature encoding outperforms the baseline from the literature, in particular, when excluding fixations on the target.

Keywords: Search Target Inference · Eye Tracking · Visual Attention · Deep Learning · Intelligent User Interfaces

1 Introduction

Human gaze behavior depends on the task in which a user is currently engaged [22,4]; this provides implicit insight into the user's intentions and allows an external observer or intelligent user interface to make predictions about the ongoing activity [6,13,2,8,1]. Predicting the target of a visual search with computational models and the overt gaze signal as input, is commonly referred to as search target inference [3,15,16]. Inferring visual search targets helps to construct and improve intelligent user interfaces in many fields, e.g., robotics [9] or similar to examples in [18]. For example, it allows for a more fine-grained generation of artificial episodic memories for situation-aware assistance of mentally impaired people [19,17]. Recent works investigate algorithmic principles for search target inference on generated dot-like patterns [3], target prediction using *Bag of Visual Words* [15], and target category prediction using a combination of gaze information and CNN-based features [16].

In this work, we extend the idea of using a *Bag of Visual Words* (`BoVW`) for classifying search targets [15]: we implement a *Bag of Deep Visual Words* model (*BoDVW*), based on image representations from a pre-trained CNN, and investigate its impact on the estimation performance of search target inference (see Figure 1). First, we reproduce the results of Sattar et al. [15] by re-implementing

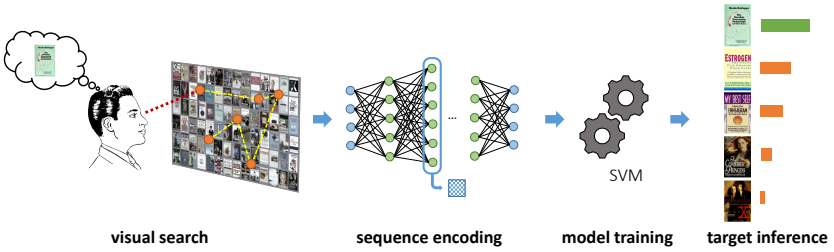


Fig. 1. Search target inference takes a fixation sequence from a visual search as input for target prediction. The pipeline we implement encodes sequences using a *Bag of Words* approach with features from a CNN for model training and inference.

their method as a baseline and evaluate our novel feature extraction approach using their published *Amazon book cover* dataset[1]. However, the baseline algorithm includes all fixations of the visual search, also the last ones that focus on the target object: the target estimation is reduced to a simpler image comparison task. Other works, including Borji et al. [3] and Zelinsky et al. [23], use fixations on non-target objects only. Consequently, we remove these fixations from the dataset and repeat our experiment with both methods. We implement and evaluate two methods for search target inference based on the *Bag of Words* feature encoding concept: (1) we re-implement the `BoVW` algorithm by Sattar et al. [15] as a baseline, and (2) we extend their method using *Bag of Deep Visual Words* (`BoDVW`) based on AlexNet.

2 Related Work

Related work include approaches for inferring targets of a visual search using the fixation signal and image-based features, as well as methods for feature extraction from CNNs.

Wolfe [20] introduces a model for visual search on images that computes an activation map based on the user task. Zelinsky et al. [23] show that objects fixated during a visual search are likely to share similarities with the target. They train a classifier using SIFT features [11] and local color histograms around fixations on distractor objects to infer the actual target. Borji et al. [3] implement algorithms to identify a certain 3×3 sub-pattern in a QR-Code-like image using a simple distance function and a voting-based ranking algorithm with fixated patches. In particular, they investigate the relation between the number of included fixations and the classification accuracy. Sattar et al. [15] consider open and closed world settings for search target inference and use the `BoVW` method to encode visual features of fixated image patches. In a follow-up work, Sattar et al. [16] combine the idea of using gaze information and CNN-based features to infer the category of a user's search target instead of a particular object instance

[1] The *Amazon book cover* dataset from Sattar et al. [15].

or image region. Similar to Sattar et al. [15], we use a *Bag of Words* for search target inference, but using deep visual words from a pre-trained CNN model.

Previous work shows that image representations from hidden layers of CNNs yield promising results for differing tasks, e.g., image clustering. Sharif et al. [12] apply CNN models for scene recognition and object detection using the L2 distance between vector representations. Donahue et al. [5] analyze how image representations generalize to label prediction, when taken from a hidden layer of a network, that was pre-trained on the ImageNet dataset [10]. We use CNN-based image features for encoding the fixation history of a visual search.

3 Visual Search Target Inference Approach

The *Bag of Words* (`BoW`) algorithm is a vectorization method for encoding sequential data to histogram representations. The `BoW` encoding is commonly used in natural language processing for, e.g., document classification [7], and was extended to a *Bag of Visual Words* for the computer vision domain for, e.g., scene classification [21]. A `BoW` is initialized with a limited set of vectors (=codewords) with a fixed size which represent distinguishable features of the data. The method for identifying suitable codewords is an essential part of the setup and influences the performance of classifiers. For encoding a sequence, each sample is assigned to the most similar codeword, resulting in a histogram over all codewords. We implement two methods based on this concept: a `BoVW` baseline similar to [15] and the CNN-based `BoDVW` encoding.

3.1 Bag of Visual Words

Sattar et al. [15] use a `BoW` approach to encode fixation sequences of visual search trials on image collages, e.g., using their publicly available *Amazon book cover* dataset that includes fixation sequences of six participants. They trained a multi-class SVM that predicts the search target from a set of five alternative covers using the encoded histories as input. We re-implement their algorithm for search target inference as a baseline including the `BoVW` encoding and the SVM target classification. Following their descriptions, we implement methods for image patch extraction from fixation sequences, a `BoVW` initialization for extracting codewords from these patches, and the histogram generation for a certain sequence. We test our algorithms using their *Amazon book cover* dataset.

3.2 Bag of Deep Visual Words

Our *Bag of Deep Visual Words* approach follows the same concept as in [15], but we encode the RGB patches using a CNN before codeword generation and mapping (see Figure 2). For this, we feed each image patch to a publicly available AlexNet model[2] which was trained using the ImageNet dataset [14] for image

[2] https://github.com/happynear/caffe-windows/tree/ms/models/bvlc_alexnet

classification. The flattened activation tensor of a particular hidden layer is used as feature vector of the input image instead of the raw RGB data. We consider the layers `conv1`, `pool2`, `conv4`, `pool5`, `fc6` and `fc8` which represent different stages of the network's layer pipeline. The patch extraction, codeword initialization (clustering) and mapping methods stay the same, but use the flattened tensor as input: the generated codewords are based on the abstract image representations of the deep CNN. Consequently, the fixation sequences get encoded using a histogram over these deep visual codewords.

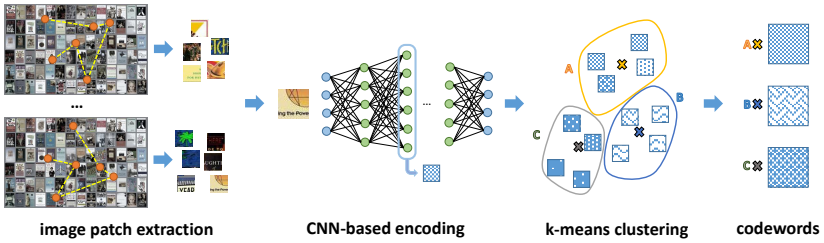


Fig. 2. For initializing the *Bag of Deep Visual Words*, image patches from fixation histories are encoded using a pre-trained CNN. The activations from a certain hidden layer are used for a k-means clustering that identifies deep codewords (cluster centers).

4 Experiment

We conduct a simulation experiment to compare the performance in predicting the search target of a visual search using our re-implementation of Sattar et al. [15]. We investigate the prediction accuracy using their `BoVW` encoding in comparison to our novel `BoDVW` encoding. We closely follow the evaluation procedure of Sattar et al. [15] for reproducing their original results using the *Amazon book cover* dataset. For this, fixations of a visual search trial are encoded for model training and target inference, also fixations on the target after it has been found. However, this is in conflict with the goal of actually inferring the search target [23,3]. Therefore, we exclude all fixations at the tail of the signal (target fixations) and repeat the experiment keeping all other parameters constant.

Sattar et al. [15] published a dataset containing eye tracking data of participants performing a search task. They arranged 84 (6×14) different book covers from Amazon in collages as visual stimuli. Six participants were asked to find a specific target cover per collage within 20 seconds after it was displayed for a maximum of 10 seconds. Fixations were recorded for 100 randomly generated collages in which the target cover appeared exactly once and was taken from a fixed set of 5 covers. Participants were asked to press a key as fast as possible after they found the target. We manually annotated each collage with a bounding box for the target cover.

In our experiment, we compare the target prediction accuracy using the `BoVW` method against our `BoDVW` encoding (using different layers). For the `BoDVW` approaches, we train multiple models, each using a different neural network layer for image patch encoding as stated in section 3.2. First, we use the *Amazon book cover* dataset with all available fixations for training and inference as proposed in [15]. Second, we repeat the experiment without the target fixations at the end of the signal. For each condition, we initialize the respective `BoW` method using a train set, encode the fixation histories (with or without target fixations) and train a support vector machine for classifying the output label.The codeword initialization and model training is performed, separate for each user (within-user condition), which yielded the best results in Sattar et al. [15]. For initializing the codewords for both approaches, we start with extracting patches around all fixations in the train set. We crop squared fixation patches with an edge length of $80px$ and generate $k = 60$ codewords. We train a One-vs-All multiclass SVM with $\lambda = 0.001$ for L1-regularization and feature normalization using Microsoft's Azure Machine Learning Studio[3]. We measure the prediction accuracy using a held-out test set as specified in Sattar et al. [15] (balanced 50/50 split per user).

We hypothesize that, using our *BoVW* implementation, we can reproduce the prediction accuracy of Sattar et al. [15] (H1.1), and that our `BoDVW` encoding improves the target prediction accuracy concerning the *Amazon book cover* dataset (H1.2). Further, we expect a severe performance drop when excluding target fixations, i.e., when using the filtered *Amazon book cover* dataset (H2.1), whereas the `BoDVW` encoding still performs better than the `BoVW` method (H2.2).

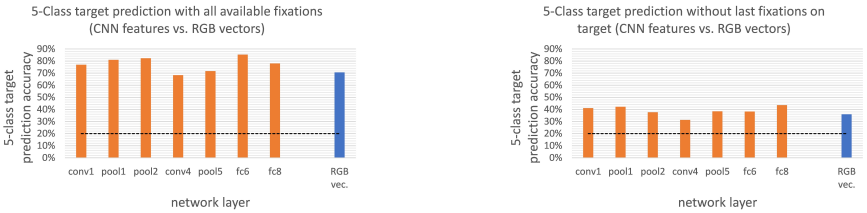
4.1 Results

Averaged over all users, our `BoVW` re-implementation of the method of Sattar et al. [15] achieved a prediction accuracy of 70.67% (20% chance) for search target inference on their *Amazon book cover* dataset with target fixations. We could reproduce their findings, even without an exhaustive parameter optimization. Concerning our *Bag of Deep Visual Words* encoding, applied in the same setting, we observe higher accuracies for all layers. The `fc6` layer performed best with an accuracy of 85.33% (see Figure 3a) which is 14.66% better compared to the baseline. When excluding the target fixations at the tail of the visual search history, the prediction accuracy of both approaches decreases: the `BoVW` implementation achieves an accuracy of 35.96% and our novel `BoDVW` encoding achieves a prediction accuracy of 43.56% using the `fc8` layer. In this setting, the `fc8` layer yields better results than the `fc6` layer with 38.26% (see Figure 3b).

5 Discussion

Our implementation of the `BoVW`-based search target inference algorithm introduced by Sattar et al. [15] achieves, with a prediction accuracy of 70.67%, a

[3] https://studio.azureml.net



(a) all fixations

(b) filtered target fixations

Fig. 3. Search target inference accuracy of 5-class SVM models using the `BoDVW` encoding with different layers (orange) and the `BoVW` encoding (blue) on (a) complete fixation sequences or (b) filtered fixation sequences.

comparable performance than stated by the authors, for the same settings (confirms H1.1). Our novel `BoDVW` encoding achieves an improvement of 14.66% with the `fc6` layer: an SVM can better distinguish between classes when using CNN features which suggests that H1.2 is correct. In the second part of our experiment, we observed a severe drop in prediction accuracy for both approaches (confirms H2.1). A probable reason is that fixation patches at the end of the search history which show the target object have a vast impact on the prediction performance: the task is simplified to an image comparison. The RGB-based codewords still enable a prediction accuracy above the chance level (20%). Our `BoDVW` approach performs 7.6% better than this baseline with the `fc6` layer (improvement of 21.13%) which suggests that H2.2 is correct. Excluding the target fixations is of particular importance for investigating methods for search target inference due to the introduced bias, hence, the procedure and results of the second part of our experiment should be used as reference for future investigations.

6 Conclusion

We introduced the *Bag of Deep Visual Words* method for integrating learned features for image classification in the popular *Bag of Words* sequence encoding algorithm for the purpose of search target inference. An evaluation showed that our approach performs better than similar approaches from the literature [15], in particular, when excluding fixations on the visual search target. The methods implemented in this work can be used to build intelligent assistance systems by augmenting artificial episodic memories with more specific information about the user's visual attention than possible before [19].

7 Acknowledgement

This work was funded by the Federal Ministry of Education and Research (BMBF) under grant number 16SV7768 in the Interakt project.

References

1. Akkil, D., Isokoski, P.: Gaze Augmentation in Egocentric Video Improves Awareness of Intention. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 1573–1584. ACM Press (2016). https://doi.org/10.1145/2858036.2858127, http://dl.acm.org/citation.cfm?doid=2858036.2858127
2. Bader, T., Beyerer, J.: Natural Gaze Behavior as Input Modality for Human-Computer Interaction. In: Eye Gaze in Intelligent User Interfaces, pp. 161–183. Springer London, London (2013). https://doi.org/10.1007/978-1-4471-4784-8_9, http://link.springer.com/10.1007/978-1-4471-4784-8{_}9
3. Borji, A., Lennartz, A., Pomplun, M.: What do eyes reveal about the mind?. Algorithmic inference of search targets from fixations. Neurocomputing **149**(PB), 788–799 (2015). https://doi.org/10.1016/j.neucom.2014.07.055, http://dx.doi.org/10.1016/j.neucom.2014.07.055
4. DeAngelus, M., Pelz, J.B.: Top-down control of eye movements: Yarbus revisited. Visual Cognition **17**(6-7), 790–811 (2009). https://doi.org/10.1080/13506280902793843, http://dx.doi.org/10.1080/13506280902793843
5. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. Icml **32**, 647–655 (2014), http://arxiv.org/abs/1310.1531
6. Flanagan, J.R., Johansson, R.S.: Action plans used in action observation. Nature **424**(6950), 769–771 (aug 2003). https://doi.org/10.1038/nature01861, http://www.nature.com/doifinder/10.1038/nature01861
7. Goldberg, Y.: Neural network methods for natural language processing. Synthesis Lectures on Human Language Technologies **10**(1), 1–309 (2017)
8. Gredeback, G., Falck-Ytter, T.: Eye Movements During Action Observation. Perspectives on Psychological Science **10**(5), 591–598 (sep 2015). https://doi.org/10.1177/1745691615589103, http://pps.sagepub.com/lookup/doi/10.1177/1745691615589103
9. Huang, C.M., Mutlu, B.: Anticipatory robot control for efficient human-robot collaboration. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 83–90. IEEE (mar 2016). https://doi.org/10.1109/HRI.2016.7451737, http://ieeexplore.ieee.org/document/7451737/
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. pp. 1097–1105. NIPS'12, Curran Associates Inc., USA (2012), http://dl.acm.org/citation.cfm?id=2999134.2999257
11. Lowe, D.: Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision pp. 1150–1157 vol.2 (1999). https://doi.org/10.1109/ICCV.1999.790410, http://ieeexplore.ieee.org/document/790410/
12. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN Features off-the-shelf : an Astounding Baseline for Recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 512–519 (2014). https://doi.org/10.1109/CVPRW.2014.131, http://arxiv.org/abs/1403.6382

13. Rotman, G., Troje, N.F., Johansson, R.S., Flanagan, J.R.: Eye movements when observing predictable and unpredictable actions. Journal of neurophysiology **96**(3), 1358–69 (sep 2006). https://doi.org/10.1152/jn.00227.2006, http://www.ncbi.nlm.nih.gov/pubmed/16687620
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
15. Sattar, H., Müller, S., Fritz, M., Bulling, A.: Prediction of search targets from fixations in open-world settings. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 981–990 (June 2015). https://doi.org/10.1109/CVPR.2015.7298700
16. Sattar, H., Bulling, A., Fritz, M.: Predicting the Category and Attributes of Visual Search Targets Using Deep Gaze Pooling (2016), http://arxiv.org/abs/1611.10162
17. Sonntag, D.: Kognit: Intelligent cognitive enhancement technology by cognitive models and mixed reality for dementia patients. In: AAAI Fall Symposium Series (2015), https://www.aaai.org/ocs/index.php/FSS/FSS15/paper/view/11702
18. Sonntag, D.: Intelligent user interfaces - A tutorial. CoRR **abs/1702.05250** (2017), http://arxiv.org/abs/1702.05250
19. Toyama, T., Sonntag, D.: Towards episodic memory support for dementia patients by recognizing objects, faces and text in eye gaze. In: KI 2015: Advances in Artificial Intelligence. vol. 9324, pp. 316–323 (2015). https://doi.org/10.1007/978-3-319-24489-1, https://link.springer.com/chapter/10.1007/978-3-319-24489-1{_}29
20. Wolfe, J.M.: Guided search 2.0 a revised model of visual search. Psychonomic Bulletin & Review **1**(2), 202–238 (Jun 1994). https://doi.org/10.3758/BF03200774, https://doi.org/10.3758/BF03200774
21. Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W.: Evaluating bag-of-visual-words representations in scene classification. In: Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval. pp. 197–206. MIR '07, ACM, New York, NY, USA (2007). https://doi.org/10.1145/1290082.1290111, http://doi.acm.org/10.1145/1290082.1290111
22. Yarbus, A.L.: Eye movements and vision. Neuropsychologia **6**(4), 222 (1967). https://doi.org/10.1016/0028-3932(68)90012-2
23. Zelinsky, G.J., Peng, Y., Samaras, D.: Eye can read your mind: decoding gaze fixations to reveal categorical search targets. Journal of vision **13**(14) (dec 2013). https://doi.org/10.1167/13.14.10, http://www.ncbi.nlm.nih.gov/pubmed/24338446http://www.ncbi.nlm.nih.gov/pubmed/24338446