Advances in Cognitive Systems X (2015) 1–6 Submitted 2/2015; published 5/2015

From Deep Learning to Episodic Memories: Creating Categories of Visual Experiences

Jigar Doshi JDOSHI8@GATECH.EDU
Zsolt Kira ZSOLT.KIRA@GTRI.GATECH.EDU
Alan Wagner ALAN.WAGNER@GTRI.GATECH.EDU
Georgia Tech Research Institute, Institute of Robotics and Intelligent Machines, Georgia Institute of Technology, Atlanta, GA 30332 USA

Abstract

This paper presents a cognitively inspired approach for visual scene categorization and abstraction. Our approach uses first-person video from real, dynamic environments to create episode-like memories of video scenes. Videos from newly encountered environments can then be matched to previous episodes and used for prediction. Our process utilizes the final layer of a convolutional neural network (CNN) as a high-level, scene-specific representation which is robust enough to noise to be used with wearable cameras. Inspired by results from cognitive science and neuroscience, we use output maps created by a CNN as a sparse, abstract representation of visual images. These output maps characterize a visual scene in terms of the spatial and temporal distribution of objects in the scene. The system is demonstrated on video taken using Google Glass. When compared with human evaluations the system correctly matches 70% of scene segments. Finally, high-level scene prediction is demonstrated by showing that the system can match scenes using only a few initial segments and can then predict the objects that will appear in the near future. Empirically, object predictions based on the initial scene segments resulted in a 95% match.

1. Introduction

This paper presents an interdisciplinary approach to the problem of scene recognition, categorization, and prediction. Our approach is inspired by neuroscience and cognitive science theories. Our implementation of these ideas draws from recent techniques from machine learning and computer vision, yet connects to high-level symbolic reasoning about categories. Our hope is that this work will serve as a conceptual and computational bridge between low-level sensor-based scene categorization techniques and high-level methods influenced by insights from human thinking.

Inspiration for our approach stems from neuroscience and cognitive science. Recent evidence suggests that concept cells may serve as an internal, neuronal representation of external stimuli and memories (Roelfsema, 2006). Studies have demonstrated that the presentation of a familiar person's face elicits a specific pattern of neural firings (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). The pattern of cell firings has been shown to relate to specific concepts. These so called "concept cells" encode specific stimuli as a sparse network of connected firing neurons. The firing patterns of these sparse networks results in the formation of new declarative memories

(Eichenbaum, 2004; Squire, Wixted, & Clark, 2007). The encoding of concepts is hierarchically organized along the processing pathway such that recordings taken from deeper layers of neurons exhibit increased abstraction indicated by their response to more complex features and greater visual invariance (Palmeri & Gauthier, 2004). This increasing abstraction culminates with hippocampal processing of abstract, personally relevant concepts into episodic memories (Logothetis & Sheinberg, 1996).

The creation of episodic memories constitutes a fast, specific method of learning. Norman and O'Reilly (2003) present a framework which combines the creation of episodic memories in the hippocampus with the extraction of generalities in the neocortex. Their framework is based on the idea that learning specific events and generalizing from those events are computationally incompatible tasks. For this reason, humans have evolved complimentary systems which allow one to memorize particular experiences quickly. Over time, separate experiences in memory are integrated to extract experiential generalities. At a low level, the researchers describe the underpinnings of their framework in terms of activation patterns of networks of neurons occurring in either the hippocampus or the neocortex. At a high level, the learning that occurs relates closely to exemplar/prototype models of category learning (Rosch, 1973; Smith & Zarate, 1990). Exemplar models of learning assume that categorical information is represented in terms of specific examples that have been experienced. Prototype models, on the other hand, categorize new stimuli with respect to an averaged representation of the category.

In previous work we approached the problem of creating exemplar models of particular experiences from a purely symbolic perspective by manually requiring people to enter symbolic descriptions of the high-level features they saw in static images (Wagner & Doshi, 2013). Results from these preliminary experiments supported the overarching idea that exemplars and prototypes could be used to create distinct categories of situations which could, in turn, then be used to predict aspects of a newly encountered scene given only the new scene's perceptual features. Unfortunately, this process was slow and not scalable to real-world problems.

Motivated by these results we sought to develop a computational process that could bridge the notions of concept cells and category learning in a manner that was implementable on a computer or robot. Recent progress has been made using deep learning to create multi-layer convolutional neural networks (Russakovsky et al., 2014). Convolutional neural networks (CNN) learn a hierarchy of visual features ranging from low-level filters to object parts to entire objects themselves. The most abstract of these features tend to display greater visual invariance and robustness, just like their neural counterparts. We reasoned that the use of output maps taken from these higher level features would allow us to categorize scenes from blurry video taken in natural environments. Intuitively, CNNs transform low-level video images into high-level, scene specific representations which are resistant to noise. As our experiments demonstrate, the system is robust enough to be used on video captured by Google Glass. Most traditional approaches to scene categorization relies on static images and a predefined number of categories (Quattoni & Torralba, 2009). Our system also does not require predefined categories.

The primary contribution of this paper is the development and demonstration of a method which characterizes scenes in terms of the presence of higher-level objects and their spatial arrangement in order to be able to match and make inferences about the objects one will see in the near future. This paper also illustrates how the raw output from a higher layer of a CNN can be used to create exemplars of specific scenes captured as video from a first-person perspective.

The remainder of the paper begins by describing our process for creating high-level concept maps from first-person video using a CNN. Next, the process for creating episode-like

representations from these concept maps is detailed. A series of experiments examining the system's ability to match video segments both in terms of the evaluations made by humans and in terms of specific categories are then described. A final experiment explores the system's ability to make predictions about the near-term appearance of visual objects in the environment based on previously experienced scenes. The paper concludes with a discussion of future work and research directions.

2. Creating High-level Concept Maps from First-Person Video

The first contribution of this work is the development of a process which allows one to create high-level representations from first-person video and to use these representations to evaluate the similarity of different scenes. A key point is that this process allows us to evaluate the distance between two frames of video not in terms of low-level representations such as pixels or edges, but rather from the perspective of high-level objects and their position in the environment.

The process begins when a CNN is used to convert individual video frames into a set of 256 output maps (Figure 1). CNNs are a class of deep learning architectures that alternate between two stages: 1) convolution, in which the inputs are convolved with learned filters that are then fed through a non-linear function and 2) pooling, which summarizes the output of a local group of output neurons. This technique was first popularized by (Lecun, Bottou, Bengio, & Haffner, 1998), who reported state-of-the-art performance in text recognition and has received significant attention recently due to achieving substantially higher performance than existing techniques in a variety of tasks including object classification, speech recognition (Krizhevsky, Sutskever, & Hinton, 2012), and text analysis (Simard, Steinkraus, & Platt, 2003). Within the computer vision field, the use of CNN's has resulted in highly accurate recognition of objects from still images.

For this paper, we use Caffe, an open-source framework for deep learning. Caffe includes several neural networks which have been pre-trained on the ImageNet dataset (Jia et al., 2014). AlexNet, the model used in this work, was trained on 1.4 million images capturing approximately 1000 different categories of object. AlexNet is capable of recognizing a thousand different objects with over 90% accuracy (Russakovsky et al., 2014) and consists of a five convolutional layer architecture with three fully connected layers.

Our research uses the output generated by the fifth convolutional layer as a higher-level representation roughly capturing the objects and their location in a scene. The output generated by the fifth layer consists of 256, 13x13 "maps." In contrast to the output from lower layers, these fifth layer output maps constitute a higher-level representation roughly capturing the objects and their location in a scene. Yet, unlike the output from higher layers, the output from the fifth layer has not been reduced to symbolic labels. The output maps from the fifth convolutional layer of the network capture the identity, strength, and spatial distribution of objects throughout the image.

The management of 256, 13x13 maps, however, is unwieldy. For this reason we use an Improved Fisher Vector in conjunction with the pre-trained Gaussian Mixture Model (GMM) to produce a fixed length encoding of the frame which summarizes the strength of association between the set of output maps to the different modes in the GMM. The resulting Fisher vector is an extremely sparse representation of the distribution of objects and their location in the image frame. The cosine distance, $\cos(FV^A, FV^B) > \alpha$, can then be used to evaluate the distance from one Fisher vector to the next. The total distance between frames is calculated by summing these individual distances.

The result of this process is an overall distance between two frames of video. Because the representation is based on high-level features such as objects and their position in the environment, we hypothesized that our approach would be robust enough to noise and blur that it could be used on video captured by a wearable camera. This is a critical distinction of the

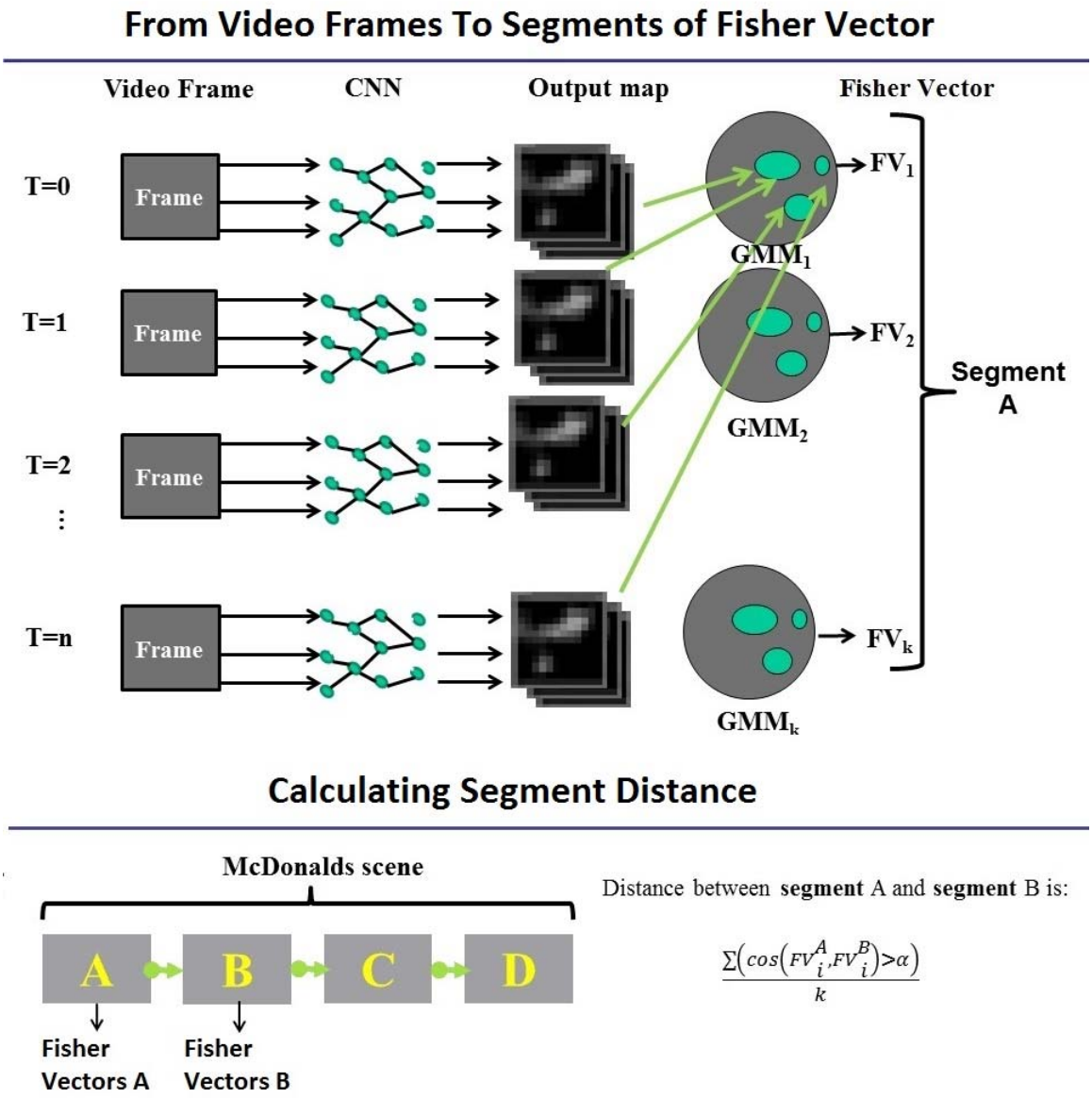


Figure 1. Our process for calculating the distances between segments of video is presented above. First, a convolutional neural network converts video frames into output maps which can roughly capture the high-level objects and their spatial arrangement in the image. Next, each output map is translated into a Fisher vector, a sparse representation of the information. A collection of Fisher vectors represents a segment of video frames. The cosine distance is used to calculate the distance between Fisher vectors.

difference between our method and other methods. We also believed that this process would allow us to cluster visual scenes with respect to higher, more abstract concepts. Our method characterizes two locations as similar if similar items are identified at the locations. For example, a McDonald's restaurant will be identified as more similar to a picnic than a library because of the objects co-identified with picnics and restaurants (i.e. utensils, glasses, plates, etc.). This leads to interesting situational characterizations when used in conjunction with first-person streaming video. For instance, while walking to a library the camera captured several images of a Starbucks coffee shop and identified the location as similar to a restaurant only while the camera viewed the coffee shop location.

3. From Concept Maps to Visual Episodes

When our process is applied to first-person video, the output maps that result bear some relation to the common notion of episodic memory. Notably, output maps: 1) contain a record of sensory processes (although no affective component is included); 2) represent visual images of the stimulus; 3) retain a first-person perspective; 4) capture short time slices of the experience; 5) capture temporal information about experiences; 6) are rapidly forgotten as individual frames are compressed into Fisher vectors representing larger segments; and 7) are autobiographical. We consider the 3-7 minute first-person videos that we captured to constitute a single episode.

Early experiments using the process indicated the similarity of one episode to another varied significantly over the course of a video. For example, a restaurant's parking lot is a very different visual environment than the establishment's bathroom. We therefore arbitrarily divided the videos into 20 seconds segments. Because they occur over a relatively short time span, these segments were meant to capture comparatively similar aspects of the episode. Moreover, segmenting compresses the episode into in series of distinct, temporally connected portions. In this manner, entire scenes can be broken down into individual segments and using the process outlined above (Figure 1), distances between segments can be generated allowing for clustering or similarity matching. Currently the segments are uniformly distributed across the episode. We are developing a technique for automatically creating segments of variable lengths.

We hypothesized that videos from the same category of environments would result in similar patterns of segments. For example, at a fast-food restaurant customers typically enter the establishment, order food, fill drinks and wait for their orders to be filled, eat, and finally leave. In other words, a restaurant episode often affords a regular pattern of segments that can be analyzed, described, compared, and used for prediction. Such regularity allows for additional information beyond objects and their spatial relationships to be used when performing tasks such as scene classification and scene abstraction. More importantly, this analysis could also allow one to predict the objects and scene fragments that will follow, an important task for several applications such as robotics. Further, even if the pattern is not strictly followed (for instance the person goes to the bathroom) the total similarity of the entire episode is likely to be greater than to episodes from dissimilar categories. The experiments that follow examine the process and its potential for making predictions about the near-term visual environments.

4. Experiments

Our experiments first required the generation of first-person video data from a variety of different environments. Recently, it has become possible to collect video from a first-person perspective using wearable cameras such as Google Glass. These videos capture one's movement, views, and

activities from a first-person perspective. Moreover, video information collected via Google Glass could potentially be used for robotics or human computer interaction tasks. The videos that we recorded captured the interactions and objects that occur in several different scenes. This platform was chosen mainly because of its small, unobtrusive design which allows a high degree of natural interaction. The Google Glass camera operates at a resolution of 1280 x 720 and a frame rate of 30 fps. The videos that resulted were not altered or preprocessed to improve quality. Nine different videos between 3-7 minutes long were recorded by two different experimenters acting independently and at different times and days. The experimenters acted naturally as they interacted with people and objects in the scene. The scenes were intentionally chosen from different categories of locations, such as fast-food restaurants, parks, and libraries. *Table 1* describes each of the nine scenes. Since the initial development of this paper we have expanded this dataset to include 34 different videos. The results present here, however, are based on these original 9 videos. Each of the videos was then divided into 20 second segments. Twenty second segments were used because it was felt that this length was long enough to capture a meaningful snapshot of a particular scene.

Table 1. Different types of recorded scenes, their category, and number of different instances.

Category	Scene Type	Number of Scenes
Fast Food	Burger King, McDonalds, Krystal	3
Library	Georgia Tech (x2), Emory	3
Park	University Park, Public Park	2
Cafeteria	Hospital Cafeteria	1

Once the videos were collected, they served as data for the experiments described below. In order to reduce the processing time, we only selected the first frame for every second of video. Each frame was then passed as input to the process described in Figure 1.

4.1 Comparisons to Human Evaluations

Our first experiment established ground truth by comparing the system's estimate of scene segment similarity to evaluations made by people. Because there are many different ways to judge similarity, the purpose of this experiment was to establish, to the extent possible, ground truth on which we could roughly gauge the accuracy of the system. We hypothesized that the system's estimate of similarity would strongly correlate to the estimates made by human subjects.

Crowdsourcing was used to obtain human subject evaluations of segment similarity. Crowdsourcing is a method for collecting data from a relatively large, diverse set of people (Paolacci, Chandler, & Ipeirotis, 2010). Crowdsourcing sites, like Amazon's Mechanical Turk, post potential jobs for crowdworkers, manage worker payment, and worker reputation. The use of crowdworkers offers a quick and efficient complement to traditional laboratory experiments. Moreover, the population of workers tends to be somewhat more diverse than traditional American university undergraduates. In order to ensure the best possible data, individuals were required to have a 95% acceptance rate for their past work and were only allowed to participate once. To ensure thoughtful evaluations, each worker was asked to briefly describe their rationale behind their evaluations. Participants were paid an effective hourly rate of $8.87. Approximately

10% of the surveys were rejected because of a failure to follow the instructions or accept the consent agreement. The accepted data originated from 224 different people.

The nine scenes from Table 1 were divided into 20 second segments resulting in a total of 161 different segments. Subjects were presented with one 20 second target segment and four randomly chosen segments. They were asked to rate the similarity of each randomly chosen segment to the target on a scale of (1-10) and to briefly describe their rationale. Once this task was complete they were then presented with a different target and set of different randomly chosen segments for evaluation. Each participant evaluated 2 groupings of targets and random segments. Because of limits on the number of subjects, only 64 of the possible 161 target segments were used in this experiment. The selection of target segments used in the experiment was random. Overall, each target-random segment combination was evaluated once by seven different people.

The evaluations made by the study's participants were compared to the evaluations made by our system in several different ways. As expected, for some segment-to-target comparisons, there was little or no consensus among the human raters in terms of similarity score. We arbitrarily defined high consensus segment-to-target evaluations as those in which the inter-rater standard deviation was less than 2. For these high consensus evaluations, the correlation with our system was $\rho(249) = +0.609$. The percent match was 76.3%, which is statistically significant ($p < 0.01$). If we consider both evaluations with and without consensus, we found a 70% agreement between our system and the participants' evaluations. This level of agreement is statistically significant from a random baseline, $p < 0.01$. We found a $\rho(384) = +0.498$ correlation between the similarity scores generated by our system and those of the participants. For data involving human subjects, this represents a strong, positive correlation (Hemphill, 2003) and supports our hypothesis.

These results strengthen our contention that the segment evaluations made by the proposed system correlate to those made by people. This is potentially important for applications that involve people, such as having the system identify locations that match a person's desired location or locations to avoid. For example, alerting a user when they have entered a room that it is similar to a bathroom may aid the visually impaired. These results do not, however, indicate the extent to which segments and scenes cluster around a category, such as restaurant. We therefore conducted an experiment exploring if segments clustered around abstract categories of locations, such as restaurants, libraries, and parks.

4.2 Segment Matching

Given the correlation to human evaluations, we decided to investigate whether the distances generated by the system could be used for matching different video scenes. If so, it might then be possible to use the system to match one's current scene to a previously experienced scene and to use that information to predict upcoming objects and events. Moreover, clusters of similarly matched video segments could potentially be generated to represent abstract categories of visual environments. To this end, we hypothesized that segments originating from the same general category (restaurant, park, etc.) would be more similar to each other than segments from different general categories. We believed that the system would generate clusters of segments which matched the general categories listed in *Table 1*. In order to test this hypothesis the system was used to generate distances between the 161 segments created in the experimental setup from Section 4.1.

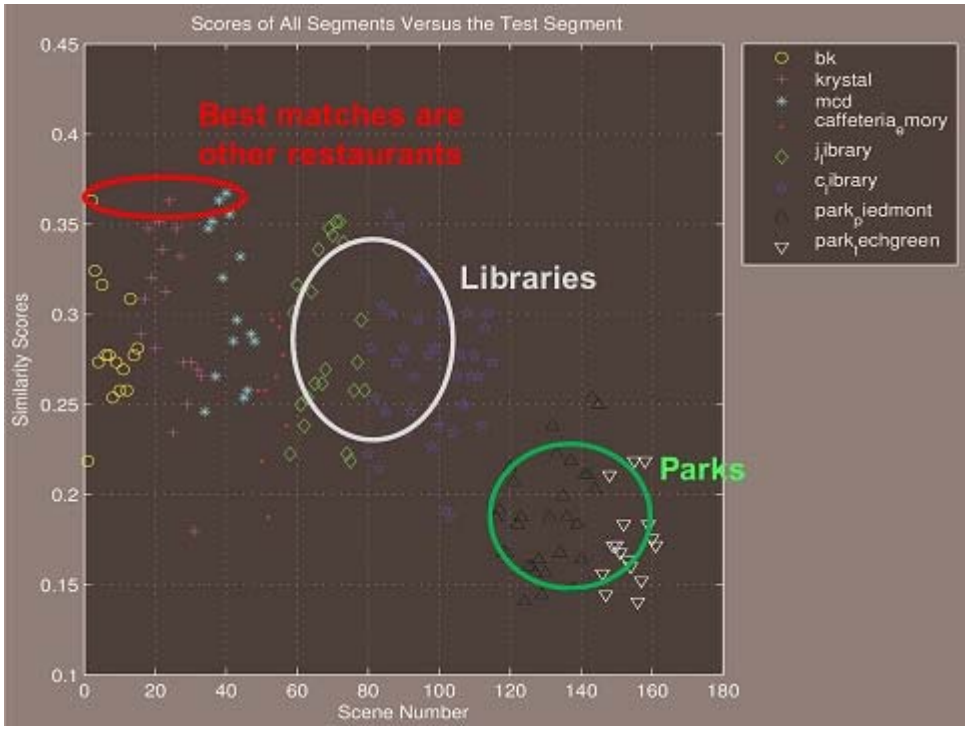


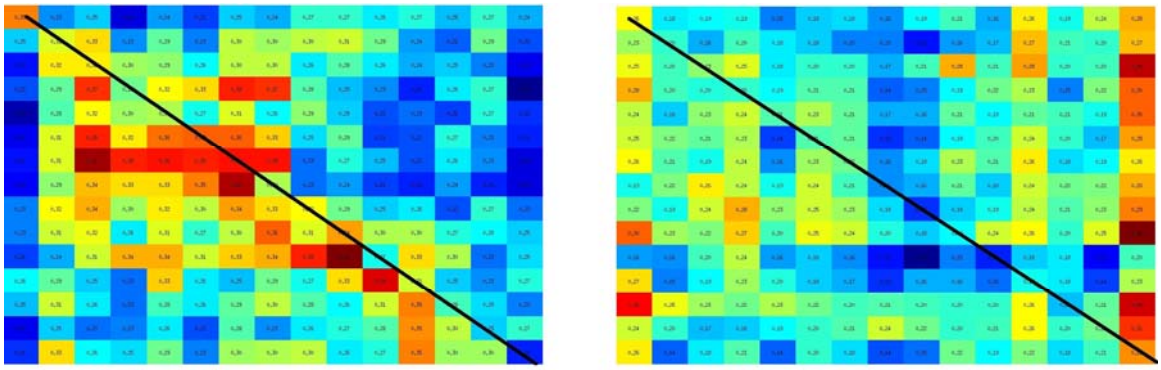
Figure 2. The graph above compares similarity scores for segments from the 8 scenes to segments taken from the McDonalds scene. The data depicts a spread across scenes with some segments matching multiple scenes. Nevertheless, the best match for each category of segments is a member of the same class. As indicated by the circle in read, the best matches to the McDonald's target are segments from the other restaurant scenes. Some segments taken from library scenes also match well to restaurant target. Finally, parks do not match well to restaurants. The individual data points correspond to different segments.

Figure 2 illustrates a single representative target segment compared to all 161 segments across all 9 scenes. The x-axis indicates the segment number and the symbol type denotes the scene of origin. The y-axis indicates the similarity score. The target segment was from the McDonalds restaurant scene. As depicted in the figure, the best matches are to the other fast food restaurants. The worst matches are to segments from park scenes. The figure also shows the skew of similarity across the scene. In other words, a range of match similarities occur with respect to the scene itself, the best of which occur in scenes from the same category. For the McDonalds scene, the best match was from a member of the restaurant category in 13 of the 15 segments (not shown in figure).

Figure 3(a) compares every segment of the Burger King scene to every segment from the McDonalds scene. The top left of the graph depicts segment 1 from each scene. The diagonal from the top left to the bottom right matches each segment temporally. Red shades indicate a closer match. The strongest red pattern occurs along the diagonal for this comparison. This indicates that the most similar segments are from temporally corresponding segments in each of these scenes. By contrast, the Figure 3(b) compares the Burger King scene to one of the Park scenes. In this case, the top left to bottom right diagonal does not indicate strong matches. Strong

matches do arise, however, in the first and last segments of the Burger King video. These matches result from video taken in the parking lot as the experimenter exits their car and enters the restaurant or leaves the restaurant. This Burger King restaurant's parking lot had several trees and bushes which were matched to the trees and bushes typically observed in the park. Overall, these results show not only can the system match specific episodes to one another, they indicate that, for matches, the visual information in the scenes is temporarily consistent. Keep in mind these scenes were recorded on different days, at different locations, and by different people.

The results support our hypothesis that scenes originating from the same abstract category generate the best matches and that the best matches correspond to similar segments within a scene. We believe that the ability to match first-person video segments to other previously experienced segments could be a powerful, new technique. Unlike the use of kernel methods popular with computer vision, our approach utilizes the CNN's ability to translate video segments



(a) McDonalds vs. Burger King (b) McDonalds vs. Piedmont Park

Figure 3. This figure is best viewed in color. It depicts a colorized version of the distances for a segment to segment comparison for complete scenes. To the left, a scene captured at McDonalds is compared to a scene captured at Burger King. The diagonal line indicates comparisons of segment (1,1), (2,2), and so on for each video. In (a) strong matches occur along this diagonal indicating that the system is evaluating the temporal pattern between these environments as similar in spite of the fact that the videos were taken by different people and at different locations and times. Moreover, the total match strength over the whole comparison is greater for (a) than (b). In (b) we see little matching along the diagonal which indicates that the temporal order of the segments taken from a restaurant are not similar to those taken at a park. There is one interesting exception, however. The first and last column in (b) represents video segments taken outdoors while the person moves from their car through the parking lot to the restaurant door. These video segments record the presence of trees, bushes, and grass which matches the objects found in a park. For these figures the park scene was clipped to 15 segments.

into a series of higher-level objects and their spatial location and uses this information to generate matches. The use of an abstract representation offers a method for categorical inference from visual input. One objective of this work is to be able to predict the high-level visual features of a soon to be encountered environment based on an individual's current environment and previous experience. The next experiment describes promising preliminary results in this area.

4.3 Episode Matching and Prediction

A system that can match its current scene to previously encountered scenes and predict which objects it will encounter in the near future would be an important step towards creating systems that can use their previous visual experiences to inform future plans. For example, the system could match a person's current video segment in which they are ordering food at a counter, to previous experiences (video scenes) at other restaurants. This information could then be used to predict the objects that will be seen in the near future (i.e. forks, spoons, etc.). Ideally this information could be used to assist the user in some way.

With this goal in mind, an experiment was conducted that examined if the initial segments from a scene could be used to match the person's current environment to a previously encountered episode. If a high percentage of segments can be matched to the correct category then this would serve as preliminary evidence that such a system might be possible.

The experiment followed the same general setup as before (Section 4.1). In this case, however, one target scene from Table 1 was withheld to represent the current scene. The system used the first x segments from the current scene to determine if the scene matched the fast food, cafeteria, or park categories (the cafeteria scene was dropped because it consisted of only a single video). Similarity scores were calculated from the current scene to all of the segments from these 8

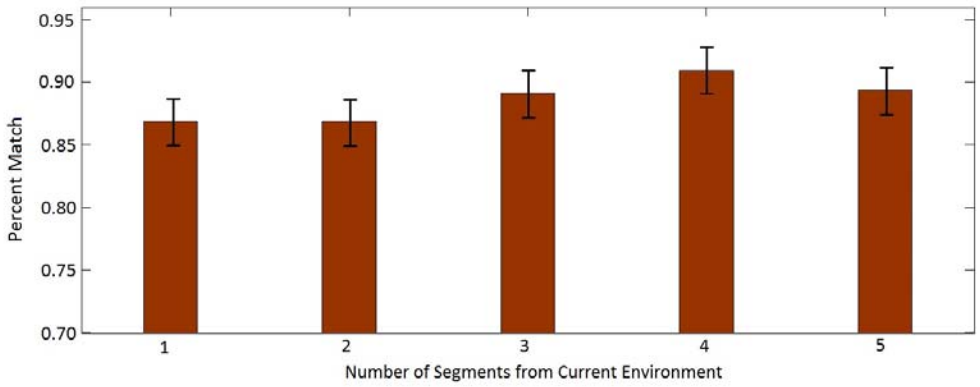


Figure 4. On average 88.6% of segments from an arbitrary current scene were matched to scenes taken from the same category. The percent match varied only slightly with respect to the number of segments from the current scene that were used to locate a match.

scenes. The match was considered correct if the most similar scene was from the same category as the current scene.

We hypothesized that the initial segments from a scene could be used to accurately match the scene to a previously experienced scene category. Initially it we believed that the accuracy of the match would depend on the number of segments used to probe memory. For example, we believed that three consecutive 20 second segments of video from the user's current environment would provide better matches than two consecutive segments.

Figure 4 depicts the results from this experiment. Importantly, the percent match is high across all cases and far exceeds a random baseline of 19.4%. The high percent match is encouraging, yet further experiments must be conducted to examine whether this result holds when a greater number of scenes are used. The results also show that as the number of segments used to select a match increases, the percent match initially increases slightly from 87.9% to 90.9% but then drops to 89.1%. The number of segments used to locate a match therefore had little impact on

performance. It is not clear to at this time why increasing the number of segments did not increase performance.

In a final experiment we evaluated the system's ability to predict future objects from a currently experienced scene (Figure 5). Here again one target scene from Table 1 was withheld to represent the current scene. The system used only a single segment from the scene to select the closest matching scene. Once a match was found, we iterated along both the current scene and its closest match, counting the number of objects predicted from the matched scene that occur in the current scene. Recall that the CNN can be used to recognize over 1000 different objects. When averaged over all segments we found a $94.76\% \pm 0.18$ correct prediction rate. This strong result supports our contention that, at least for the setup tested, the system can use experiences from the past to predict which objects an individual will encounter in the future.

Overall, the results from these three experiments are meant to convey the breadth of what we expect this approach to achieve. The results hint that the system can be used for prediction and to match one's experiences in a particular scene to a category of scenes and previous experiences. Moreover, because the experiments used raw data taken from a wearable computer, we have reason to believe that practical applications derived from the system are achievable in the near-term. One area for future work will focus on strengthening the results by expanding the set of scenes considered.

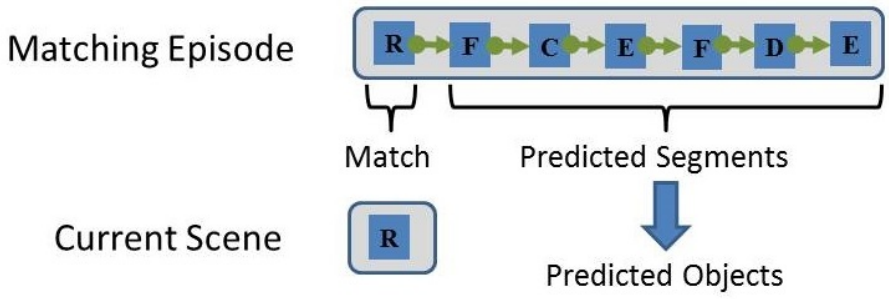


Figure 5. The final experiment first matches segments from the current scene to a previously experienced episode. The matched episode's remaining segments are used to predict which objects the individual will encounter in the near future.

5. Related Work

Implementations of episodic memory have been investigated previously as part of several different cognitive architectures. Soar is a well-established architecture that includes a model of human working memory (Anderson et al., 2004; Laird, Newell, & Rosenbloom, 1987; Langley & Cummings, 2004; Winkler, Tenorth, Bozcuoglu, & Beetz, 2014). Soar's model of memory, for instance, includes representations of procedural, declarative, and episodic memory. Although the different architectures are unique and offer several different tradeoffs, each relies on symbolic representations of memories. In contrast, the system that we present relies on connectionist representations of high-level information as a sparse network of neural-network based activations. From an architectural standpoint, we envision the process described in this paper serving as a

bridge connecting low-level sensor input to higher level symbolic representations to assist with human-level reasoning.

Computational models of visual analogy have been developed and tested on visual intelligence tests (Davies, Goel, & Yaner, n.d.; Gentner, 1983). These systems tend to first translate visual content into symbolic representations and then use analogical reasoning on the symbolic representation. Our approach differs in that it does not employ a symbolic representation but can be converted to a symbolic system in order to tie-in with these more developed systems.

Research in the area of scene classification and recognition has recently become an important field of focus. Researchers have predominately focused on the task of classifying single images in terms of a predefined labeled category (Fei-Fei & Perona, 2005; Juneja, Vedaldi, Jawahar, & Zisserman, 2013; Oliva & Torralba, 2001). Traditionally, these approaches tend to use low-level feature descriptors such as SIFT (Lowe, 2004) or HOG (Dalal & Triggs, 2005) to characterize the target scene. Xiao et al. (2010), for example, utilized an extensive dataset of different scenes derived from over 130,000 images to test a variety of low-level feature descriptors on the task of scene classification. Other systems use predefined categories of scene images and learn classifiers with respect to these categories (Quattoni & Torralba, 2009). In contrast, our process does not require predefined categories and matches streaming video, a more challenging dynamic problem.

More recently, CNNs have been used in which a variety of features are learned (Girshick, Donahue, Darrell, & Malik, 2013; Krizhevsky et al., 2012; Lecun et al., 1998). Donahue et al. (2013), for instance used deep CNNs to examine the task of scene classification. They used Caffe to classify scenes from the SUN-397 scene category database obtaining a recognition rate of 40.94%. Later work by the same group explored the use of long short term recurrent convolutional networks on video sequence classification (Donahue et al., 2014). Recurrent neural networks have recently become a popular method for learning representations of video (Mao et al., 2014; Srivastava, Mansimov, & Salakhutdinov, 2015; Venugopalan et al., 2014).

For several reasons, we have intentionally avoided directly comparing our system to these systems. First, the purpose of our system is different. Our motivation is not simply to create a system which is state-of-the-art in scene classification. Rather, our purpose is to develop a process that generates perception grounded representations which can be used for higher-level reasoning. Second, our approach was designed for video generated by a wearable camera. Most approaches by the deep learning community rely on datasets that have been neatly partitioned into precise action categories, such as the UCF-101 dataset. Although we did not show it, the videos created by a wearable camera would likely not be well classified by neural network trained on action videos. Finally, many of the training and test videos used in this research are from a stationary camera. Because one primary goal of this research is to implement this system on a moving first-person camera, we felt that it was inappropriate to compare this work to video and images taken from a stationary camera included within these datasets (Xiao et al., 2010). Our approach differs in that we utilize the output maps generated by higher layers of the CNN. The more abstract features captured in these higher layers are relatively stable across the frames of a video in spite of blur and camera motion. This facet of the research makes our system implementable with a wearable camera.

6. Conclusion

This paper has presented a system that uses the output maps generated from a CNN with first-person camera video input to create concept cell inspired representation of the objects and their

spatial location in a scene. We have shown that methods for calculating the distance between the output maps generated by a CNN afford a means for determining scene similarity and/or distance. We have demonstrated that: 1) segment similarity correlates to the similarity judgments of people; 2) that if the scene is broken into segments then these similarity scores can be used to cluster scenes with respect to abstract categories and; 3) that the system offers a potential means for predicting future encounters with objects.

The connection of the system we present to cognitive systems is worth highlighting. Although our system does not include traditional high-level reasoning functionality, it does connect real world perception in unadulterated environments to the precursors of higher level reasoning, namely categories of episodic memories. Several established cognitive architectures have also used real-world perception (e.g. Robo-SOAR and ACT-R/E) (Laird, Yager, Hucka, & Tuck, 1991; Trafton et al., 2013). Unlike these approaches, the perceptual information our system uses is completely unstructured and ethologically valid. The episodic memories fashioned from our process could potentially be used to reason about novel environments, perform analogical reasoning, or perhaps even visualize an environment by recalling or restructuring the images on which the output maps are based. Overall, we feel that this process could be used as a tool to supply higher-level cognitive architectures with structured, experiential information from unstructured video.

Admittedly, our current system has several limitations. One limitation is that it takes many hours to process additional scenes for matching. The time to find a match for a scene, on the other hand, is approximately 2-4 seconds, which may be a limitation for applications that require fast scene recognition. The system is not, however, very sensitive to blur or noise. Because of naturally occurring head motion, the videos we captured tended to be blurry. Yet we chose not to deblur the videos or to remove blurred frames. The video was input to the system without further processing. The fact that we tested on consumer grade hardware and that the system did not require preprocessing to reduce blur is testament to the robustness of our approach. We are currently in the process of testing our method on a larger dataset which includes 15 different scenes.

To the best of our knowledge, this work represents the first time that the output maps from a CNN have been used as a sparse first-person representation of the visual environment for the purpose of scene understanding. We believe that our approach can be used to reason about one's location from visual information, for visual planning, and for higher-level scene abstraction.

Future work will focus on creating visual prototypes and mental visualizations of future environments. We believe that visual prototypes can be created by extracting the segments which are common to a category of experiences. Mental visualizations, on the other hand, can likely be created by combining segments or even portions of output maps to create new segments and episodes. The system might also afford a new method for vision and landmark based navigation to aid a person or a robot. The possibility recognizing scenes and using experiences from similar scenes to predict the presence of objects in the near-term visual future would be a large step towards developing an artificially intelligent system.

Acknowledgements

We would like to thank the reviewers for their thoughtful and detailed comments and suggestions as well as the Office of Naval Research for funding this work under award N00141210484.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. http://doi.org/10.1037/0033-295X.111.4.1036

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005* (Vol. 1, pp. 886–893 vol. 1). http://doi.org/10.1109/CVPR.2005.177

Davies, J., Goel, A. K., & Yaner, P. W. (n.d.). *Proteus: Visual Analogy in Problem Solving Abstract*.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2014). Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *arXiv:1411.4389 [cs]*. Retrieved from http://arxiv.org/abs/1411.4389

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2013). DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *arXiv:1310.1531 [cs]*. Retrieved from http://arxiv.org/abs/1310.1531

Eichenbaum, H. (2004). Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron*, *44*(1), 109–120. http://doi.org/10.1016/j.neuron.2004.08.028

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005* (Vol. 2, pp. 524–531 vol. 2). http://doi.org/10.1109/CVPR.2005.16

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170. http://doi.org/10.1016/S0364-0213(83)80009-3

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv:1311.2524 [cs]*. Retrieved from http://arxiv.org/abs/1311.2524

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*(1), 78–79. http://doi.org/10.1037/0003-066X.58.1.78

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., … Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv:1408.5093 [cs]*. Retrieved from http://arxiv.org/abs/1408.5093

Juneja, M., Vedaldi, A., Jawahar, C. V., & Zisserman, A. (2013). Blocks That Shout: Distinctive Parts for Scene Classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 923–930). http://doi.org/10.1109/CVPR.2013.124

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, *33*(1), 1–64. http://doi.org/10.1016/0004-3702(87)90050-6

Laird, J. E., Yager, E. S., Hucka, M., & Tuck, C. M. (1991). Robo-Soar: An integration of external interaction, planning, and learning using Soar. *Robotics and Autonomous Systems*, *8*(1-2), 113–129. http://doi.org/10.1016/0921-8890(91)90017-F

Langley, P., & Cummings, K. (2004). Hierarchical skills and cognitive architectures. In *in Proc. 26th Annu.Conf. Cogn. Sci. Soc* (pp. 779–784).

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. http://doi.org/10.1109/5.726791

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621. http://doi.org/10.1146/annurev.ne.19.030196.003045

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, *60*(2), 91–110. http://doi.org/10.1023/B:VISI.0000029664.99615.94

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2014). Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *arXiv:1412.6632 [cs]*. Retrieved from http://arxiv.org/abs/1412.6632

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646. http://doi.org/10.1037/0033-295X.110.4.611

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, *42*(3), 145–175. http://doi.org/10.1023/A:1011139631724

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews. Neuroscience*, *5*(4), 291–303. http://doi.org/10.1038/nrn1364

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). *Running Experiments on Amazon Mechanical Turk* (SSRN Scholarly Paper No. ID 1626226). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1626226

Quattoni, A., & Torralba, A. (2009). Recognizing indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009* (pp. 413–420). http://doi.org/10.1109/CVPR.2009.5206537

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107. http://doi.org/10.1038/nature03687

Roelfsema, P. R. (2006). Cortical Algorithms for Perceptual Grouping. *Annual Review of Neuroscience*, *29*(1), 203–227. http://doi.org/10.1146/annurev.neuro.29.051605.112939

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. http://doi.org/10.1016/0010-0285(73)90017-0

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge. *arXiv:1409.0575 [cs]*. Retrieved from http://arxiv.org/abs/1409.0575

Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on*

Document Analysis and Recognition, 2003. Proceedings (pp. 958–963). http://doi.org/10.1109/ICDAR.2003.1227801

Smith, E. R., & Zarate, M. A. (1990). Exemplar and Prototype Use in Social Categorization. *Social Cognition*, *8*(3), 243–262. http://doi.org/10.1521/soco.1990.8.3.243

Squire, L. R., Wixted, J. T., & Clark, R. E. (2007). Recognition memory and the medial temporal lobe: a new perspective. *Nature Reviews Neuroscience*, *8*(11), 872–883. http://doi.org/10.1038/nrn2154

Srivastava, N., Mansimov, E., & Salakhutdinov, R. (2015). Unsupervised Learning of Video Representations using LSTMs. *arXiv:1502.04681 [cs]*. Retrieved from http://arxiv.org/abs/1502.04681

Trafton, G., Hiatt, L., Harrison, A., Tamborello, F., Khemlani, S., & Schultz, A. (2013). ACT-R/E: An Embodied Cognitive Architecture for Human-Robot Interaction. *Journal of Human-Robot Interaction*, *2*(1), 30–55.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R. J., & Saenko, K. (2014). Translating Videos to Natural Language Using Deep Recurrent Neural Networks. *CoRR, abs/1412.4729*. Retrieved from http://arxiv.org/abs/1412.4729

Wagner, A. R., & Doshi, J. (2013). Who, How, Where: Using Exemplars to Learn Social Concepts. In G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *Social Robotics* (pp. 481–490). Springer International Publishing. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-02675-6_48

Winkler, J., Tenorth, M., Bozcuoglu, A. K., & Beetz, M. (2014). CRAMm -- Memories for Robots Performing Everyday Manipulation Activities. *Advances in Cognitive Systems*, *3*.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3485–3492). http://doi.org/10.1109/CVPR.2010.5539970