Advances in Transportation, Vol. 60, 2022, 165–171
https://doi.org/10.54941/ahfe1002446

Visual Dictionary of Human Action in Vehicular Environment Using Computer Vision

Abhijit Sarkar

Virginia Tech Transportation Institute, Blacksburg, VA 24060, USA

ABSTRACT

Every human behavior and actions can be divided in small sub-actions and attributes. Some of these sub-actions and attributes contain visually semantic meanings to human. We call the ensemble of them as visual dictionary. The visual dictionaries help to create an action like how words from dictionary helps to create sentences. In this work, we demonstrate the effectiveness of the visual dictionary by analyzing driver behavior inside the vehicle. We take the primary driving behavior which includes two hands on the wheel and 56 secondary behaviors that include talking over handheld phone, eating sandwich, drinking from bottle, smoking, reaching for objects, and dancing. Finally, we demonstrate how each of these dictionary elements can be automatically processed from videos using computer vision.

Keywords: Secondary behavior, Visual dictionary, Cognitive load, Computer vision

INTRODUCTION

Human action is complex in nature. Human factor research often concentrates to understand how a series of human actions are completed, how human interacts with the visual scene, what is the effect of individual human actions in completing a task, what is their effect of each task on the cognitive load (Sanders and McCormick, 1998; Wickens et al, 2004). As humans can perform simultaneous actions at any point of time, this makes the analysis of individual actions and their interactions more difficult. In this paper we propose a new method to systematically understand any human action using visual dictionary. As human performs any action in a three-dimensional scene, each action can be defined by a series of visual attributes. In recent years, artificial intelligence (AI) has made enormous progress and can process a visual scene including human action and interactions. But most of the AI methods lacks explainibility. Therefore, in this paper, we introduce a set of visual attributes that are semantically explainable to humans but easily detectable buy a machine. We have specifically looked at driver's behavior inside a vehicle through the construct of secondary behaviors using the elements of visual dictionary.

Secondary behaviors are often regarded as major reasons for crashes and near crashes (Klauer et al, 2014; Farmer et al, 2015). Drivers' objective is to perform a primary task of driving which includes navigating a roadway scene

and perform appropriate longitudinal and lateral maneuver and safe interaction with other roadway agents including other vehicles, and pedestrians. However, drivers tend to engage themselves in different secondary tasks that can often distract them from their primary task of driving (Engström, Johansson, and Östlund, 2005). These distractions include visual distractions, auditory distractions, and cognitive distraction. Each secondary task contributes to different amounts of distraction to drivers (Liang, Reyes and Lee, 2007). Interestingly due to the complex construction of each of the secondary tasks and the overlap of the secondary and tertiary tasks with the primary task of driving it is very difficult to properly understand the role of each secondary task in inattention as well as cognitive overload. This paper shows how to define each secondary task buy a collection of low-level visual attributes. In computer vision research, researchers have often used visual attributes to define an object or an action. This work is partially inspired by them.

VISUAL DICTIONARY OF HUMAN ACTION (VDHA)

Any human action has multiple visual components. We can define any action by a set of words and attributes. These attributes carry semantic meaning that human can understand. The VDHA aims to capture the temporal sequence of body parts, interaction between objects and parts in a systematic way such that any human action can be uniquely constructed. As a result, any human action can be defined in an analytical framework to apply advanced data analytics techniques and machine learning algorithms for automated discovery and analysis.

Definitions: A *visual dictionary of human action* is a collection of *elements* where each *element* captures unique attribute and/or relation of different human body parts, and interactions to the *object* and *stuffs* in their surrounding environment. Each *element* can be further defined by a predetermined set of *values*. Any human action can be uniquely constructed using these elements and their values.

The elements can either be i) micro actions like movement of head, movement of eyelid etc. ii) Spatial relation between different body parts like distance between face and hands iii) Direction of movements like forward, sidewise etc. iv) interaction with objects in the surrounding world like holding, lifting etc. v) Temporal patterns like repetition of micro actions, frequency of micro actions etc. Each element can take multiple values including binary, nominal, or ordinal. All the combinations of the {element, value} pairs construct the feature space of the VDHA. It is expected that any human action can be constructed by a unique set of these {element, value} pairs.

Visual Dictionary for Secondary Behavior

In this section we demonstrate the design of visual dictionary and use them to define secondary action for driving task. We further demonstrate how these dictionaries can be used to uniquely define any behavior and semantic relations between secondary tasks. We have designed a total of eight dictionary elements to define secondary behaviors as shown in Table 1 i) Head movement, hand movement, and mouth movement reflects the micro actions;

Table 1. Examples of secondary behavior description using visual dictionary.

Secondary task name	Head movement	Hand movement	mouth movement	At least one Hand off wheel?	Body Pose off normal?	Head pose off normal	Hand to face distance	Presence of object
Cell phone, Talking	Maybe	N	Y	Y	N	N	Close	Y
Adjusting radio	Y	Y	N	Y	Maybe	Maybe	Far	N
Applying Makeup	Maybe	Y	Maybe	Y	Maybe	Maybe	Close	Y
Eating with utensils	Maybe	Y	Y	Y	N	N	Close	Y
Looking at pedestrian	Y	N	N	N	N	Y	NA	N

ii) Body pose, head pose, and face to hand distance reflects the spatial relation between different body parts; iii) Hands on wheel, and the presence of objects reflect the interaction with the surrounding world. as this visual dictionary is targeted to driving behaviors any normal attribute refers to the primary task of driving. Table 1 further shows how different secondary tasks can be uniquely defined by these eight dictionary elements. While constructing the task definition, we only assign values daughter is sensually required to perform the task. An element value "maybe" is used when the task can be performed with or without the attribute. For example, a person can adjust the radio with or without moving their head away from forward (which is normal value for primary task of driving). as we can see that each secondary task takes a unique combination of element and value, hence, giving them a unique signature.

Using Visual Dictionary to Understand Relations Between Different Secondary Behaviors

Once the dictionary elements are defined and the secondary task are constructed, we can use the element value pairs to construct feature vector for each of the secondary tasks. After that we use the feature vector and calculate distance between these vectors in the high dimensional feature space to understand the similarities and differences between secondary behaviors. In this case, we have used Euclidean distance and clustering methods to identify secondary tasks that are semantically similar to each other. We have used a total of 52 possible secondary behaviors from SHRP2 naturalistic driving study (Dingus et al, 2015). The secondary behaviors are selected by human annotators. The element-value pairs are annotated by six different annotators who are experts in human factor research and computer vision. As seen from the Figure 1 several secondary behaviors are clustered together. For example, any distraction-based behaviors are clustered together, but they are different from other secondary behaviors like eating, drinking or cell phone holding. Similarly, any task related to the instrument panel in the car is clustered together but they have differences from actions related to passenger interactions. This clearly shows how visual dictionary can be used to uniquely identify a human action, cluster similar actions together, and differentiate between actions which are semantically different.

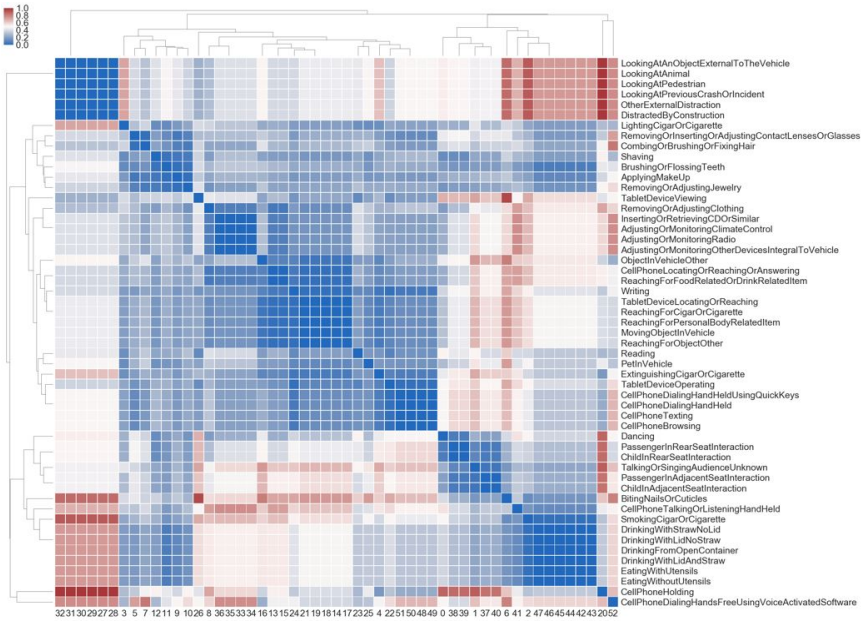
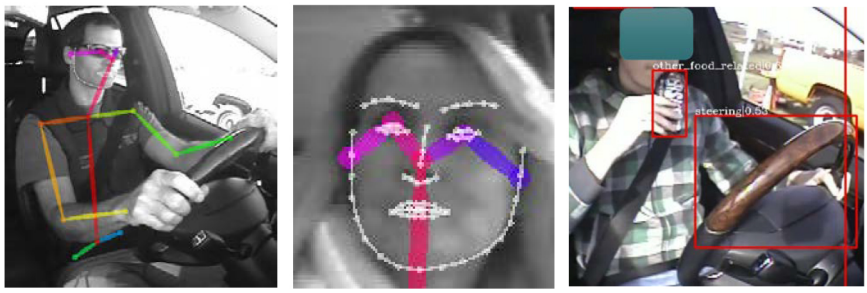


Figure 1: The relation between each of the secondary behaviors using visual dictionary. Blue represents those two actions have similar attributes (small distance), while red represents that the actions are different in nature. (Best viewed in electronic copy).



(a) (b) (c)

Figure 2: Computer vision can be used to detect and track fiducial points. We repent results from Openpose using VTTIMLP01 dataset to show how each of the fiducial points from body (a) and face (b) can be identified. (c) shows how object detectors can be used to identify the location of all the objects inside the vehicle including steering and food items.

AUTOMATIC DETECTION OF VISUAL DICTIONARY USING COMPUTER VISION

In recent years, computer vision (CV) and machine learning has made enormous advancements. Due to the development of deep convolutional neural network (Lecun, Bengio, and Hinton, 2015), we can process any image or video to accurately extract semantic elements. In this section we demonstrate how these advancements in computer vision can be used to automatically extract visual dictionary elements from images and videos. we specifically demonstrate the power of human body pose detection (Cao et al. 2017), head

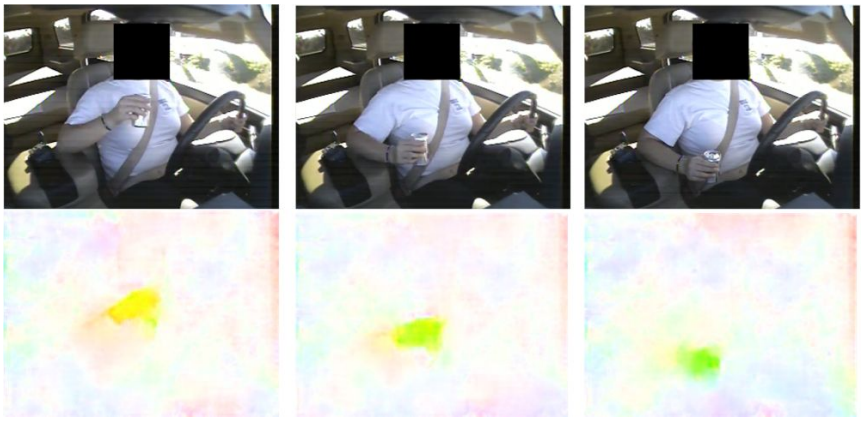


Figure 3: Optical flow-based methods to understand the movement of the fiducial points. In this example. The participant is seen to completing a 'drinking from can' task. The hand movement is downwards and away from mouth/face. Optical flow captures that movement.

Table 2. Automatic visual dictionary processing using computer vision.

	Head pose detection	Face key points	Temporal movement	Body key points	Object detection
Head movement	Y		Y		
Hand movement			Y	Y	
mouth movement		Y	Y		
At least one Hand off wheel?				Y	Y
Body Pose off normal?				Y	
Head pose off normal	Y				
Hand to face distance		Y		Y	
Presence of object					Y

pose detection, optical flow based temporal movement, and object detection (Ren et al, 2015). The body pose detection algorithm helps us to detect the pixel location of each of the joints in the body including the shoulder, elbow, neck, waist, and wrist (Figure 2 (a)). These special locations can be used to understand the relative positions between each body parts, and the posture of body. The key point locations in the face helps us to detect and track movement of different deformable parts like eyes, mouth, and jaws (Figure 2 (b)). An object detector helps us to detect all the objects that are present in the scene (Figure 2 (c)), their location and their distance from different fiducial points in the human skeleton. The temporal movement can be identified using optical flow based method which takes difference between two consecutive frames and reports movement of pixels (Baker et al, 2011). Figure 3 shows an example of hand movement of a participant who is completing "drinking from a can". the optical flow based method particularly indicates the spatial location of the movement and the speed of the movement. Using all these CV algorithms, the eight elements from secondary behavior visual dictionary can be automatically detected and located from the image on video of

a particular scene. Table 2 shows a summary of the CV tools for dictionary element processing.

CONCLUSION

In this paper, we introduce a visual dictionary to describe human action. Using these visual dictionaries, we have shown that any human action can be reconstructed and can have a unique signature. We believe that these dictionaries can be extended to other domains of applications and can be generalized. Due to its associative nature of modeling, concurrent behavior modeling is possible when a person is involved in multiple actions at the same time. This method also demonstrates a systematic way to understand task associated measures including cognitive load. Finally, we demonstrate an automated method using computer vision to automatically generate this dictionary elements and study their relations.

ACKNOWLEDGMENT

This work was supported in part by FHWA award no. 693JJ319C000004, "Video Analytics for Automatic Annotation of Driver Behavior and Driving Situations in Naturalistic Driving Data.". The opinions in this paper are those of the author and do not reflect those of any government agency. The author would like to thank Calvin Winkowski for his help in data processing. The author would also like to thank Lynn Abbott, and Jeff Hickman from VTTI for their valuable suggestions.

REFERENCES

Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J. and Szeliski, R., 2011. A database and evaluation methodology for optical flow. *International journal of computer vision*, *92*(1), pp. 1–31.

Cao, Z., Simon, T., Wei, S.E. and Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7291–7299).

Dingus, T.A., Hankey, J.M., Antin, J.F., Lee, S.E., Eichelberger, L., Stulce, K.E., McGraw, D., Perez, M. and Stowe, L., 2015. *Naturalistic driving study: Technical coordination and quality control* (No. SHRP 2 Report S2-S06-RW-1).

Engström, J., Johansson, E. and Östlund, J., 2005. Effects of visual and cognitive load in real and simulated motorway driving. Transportation research part F: traffic psychology and behaviour, 8(2), pp. 97–120.

Farmer, C.M., Klauer, S.G., McClafferty, J.A. and Guo, F., 2015. Relationship of near-crash/crash risk to time spent on a cell phone while driving. *Traffic injury prevention*, *16*(8), pp. 792–800.

Klauer, S.G., Guo, F., Simons-Morton, B.G., Ouimet, M.C., Lee, S.E. and Dingus, T.A., 2014. Distracted driving and risk of road crashes among novice and experienced drivers. *New England journal of medicine*, *370*(1), pp. 54–59.

Liang, Y., Reyes, M.L. and Lee, J.D., 2007. Real-time detection of driver cognitive distraction using support vector machines. *IEEE transactions on intelligent transportation systems*, *8*(2), pp. 340–350.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *nature*, *521*(7553), pp. 436–444.

Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

Sanders, M.S. and McCormick, E.J., 1998. Human factors in engineering and design. *Industrial Robot: An International Journal*.

Wickens, C.D., Gordon, S.E., Liu, Y. and Lee, J., 2004. *An introduction to human factors engineering* (Vol. 2). Upper Saddle River, NJ: Pearson Prentice Hall.

Yang, M., Zhang, L., Feng, X. and Zhang, D., 2014. Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, *109*(3), pp. 209–232.