

# The Action Similarity Labeling Challenge

Orit Kliper-Gross, Tal Hassner, Lior Wolf

**Abstract**—Recognizing actions in videos is rapidly becoming a topic of much research. To facilitate the development of methods for action recognition, several video collections, along with benchmark protocols, have previously been proposed. In this paper we present a novel video database, the “Action Similarity LABELiNg” (ASLAN) database, along with benchmark protocols. The ASLAN set includes thousands of videos collected from the web, in over 400 complex action classes. Our benchmark protocols focus on action *similarity* (same/not-same), rather than action classification, and testing is performed on *never-before-seen* actions. We propose this data set and benchmark as a means for gaining a more principled understanding of what makes actions different or similar, rather than learning the properties of particular action classes. We present baseline results on our benchmark, and compare them to human performance. To promote further study of action similarity techniques, we make the ASLAN database, benchmarks, and descriptor encodings publicly available to the research community.

**Index Terms**—Action recognition, Action similarity, Video database, Web videos, Benchmark.

## 1 INTRODUCTION

RECOGNIZING human actions in videos is an important problem in Computer Vision with a wide range of applications, including video retrieval, surveillance, man-machine interaction, and more. With the availability of high bandwidth communication, large storage space and affordable hardware, digital video is now everywhere. Consequently, the demand for video processing, particularly effective action recognition techniques, is rapidly growing. Unsurprisingly, action recognition has recently been the focus of much research.

Human actions are complex entities taking place over time and over different body parts. Actions are either connected to a context (e.g., swimming) or context free (e.g., walking). What constitutes an “action” is often undefined, and so the number of actions being performed is typically uncertain. Actions can vary greatly in duration; some actions being instantaneous whereas others prolonged. They can involve interactions with other people, or static objects. Finally, they may include the whole body or be limited to one limb. Figure 1 provides examples, from our database, of these variabilities.

To facilitate the development of action recognition methods, many video sets, along with benchmark protocols, have been assembled in the past. These attempt to capture the many challenges of action recognition. Some examples include the KTH [1] and Weizmann [2] databases, and the more recent, Hollywood, Hollywood2 [3], [4], and YouTube-actions databases [5].

This growing number of benchmarks and data sets is reminiscent of the data sets used for image classification and face recognition. However, there is one important difference: image sets for classification and recognition

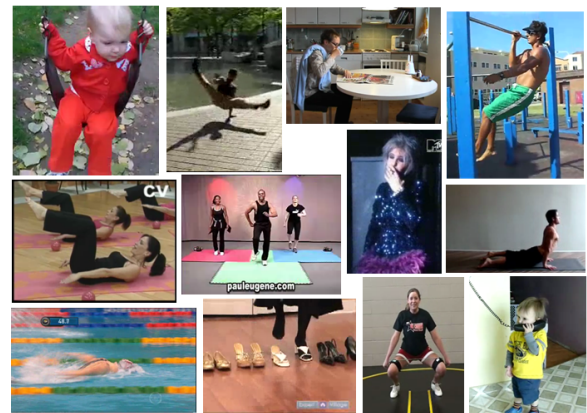


Fig. 1: Examples of the diversity of “real-world” actions

now typically contain hundreds, if not thousands, of object classes or subject identities (see for example: [6], [7], [8]), whereas existing video data sets typically provide only around 10 classes (see Section 2).

We believe one reason for this disparity between image and action classification is the following. Once many action classes are assembled, classification becomes ambiguous. Consider, for example, a high jump. Is it “running”? “jumping”? “falling”? Of course, it can be all three and possibly more. Consequently, labels assigned to such *complex* actions can be subjective and may vary from one person to the next. To avoid this problem, existing data sets for action classification offer only a small set of well-defined, *atomic* actions, which are either periodic (e.g., walking), or instantaneous (e.g. answering the phone).

In this paper we present a new action recognition data set, the “Action Similarity LABELiNg” (ASLAN) collection. This set includes thousands of videos collected from the web, in over 400 complex action classes. <sup>1</sup>

To standardize testing with this data, we provide a “same/not-same” benchmark, which addresses the action recognition problem as a non class-specific similarity

- O. Kliper-Gross is with the Department of Mathematics and Computer Science, The Weizmann Institute of Science, Israel.  
E-mail: orit.kliper@weizmann.ac.il
- T. Hassner is with The Computer Science Division, The Open University, Israel.
- L. Wolf is with the Blavatnik School of Computer Science, Tel-Aviv University, Israel.

<sup>1</sup> Our video collection, benchmarks, and related additional information is available at:  
<http://www.openu.ac.il/home/hassner/data/ASLAN/>.















tion cascade for visual identification from one example,” in *Proc. 10th IEEE Int. Conf. Comput. Vision*, vol. 1, 2005, pp. 286–293.

- [43] L. Wolf, T. Hassner, and Y. Taigman, “Descriptor based methods in the wild,” in *Faces in Real-Life Images Workshop in European Conf. Comput. Vision*, 2008.
- [44] M. Sargin, H. Aradhye, P. Moreno, and M. Zhao, “Audiovisual celebrity recognition in unconstrained web videos,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 1977–1980.
- [45] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011, available: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [46] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.



**Orit Kliper-Gross** is a PhD student at the Weizmann Institute of Science, the department of Applied Mathematics and Computer Science, the computer Vision Lab. She is working under the joint supervision of Prof. Ronen Basri and Dr. Tal Hassner. Previously, she studied in the Adi Lautman Interdisciplinary Program for Outstanding Students (studies directly for a Masters degree) at Tel-Aviv University, and received her Msc in Computer Science with honours, from Tel-Aviv University, where she worked under the

supervision of Prof. David Horn. Her research interests are in Computer Vision, including Action Recognition and Video Analysis.



**Tal Hassner** is a member of the senior faculty at the Open University of Israel since 2008. In addition, he is an adjunct faculty member at the Academic College of Tel-Aviv Yaffo. He received his BA in Computer Science at the Academic College of Tel-Aviv Yaffo in 1998. His MSc. and PhD degrees, both in Applied Mathematics and Computer Science, were received in 2002 and 2006 (respectively), from the Weizmann Institute of Science. He later completed a postdoctoral fellowship, also at the Weizmann institute. His

distinctions include the best Student Paper Award at the IEEE Shape Modeling International (SMI) Conference, 2005, and the AIM@SHAPE Best paper award 2005. His research interests are in Computer Vision, including face-recognition and single-view, 3D-reconstruction.



**Lior Wolf** is a faculty member at the School of Computer Science at Tel-Aviv University. Previously, he was a post-doctoral associate in Prof. Poggio’s lab at MIT. He graduated from the Hebrew University, Jerusalem, where he worked under the supervision of Prof. Shashua. Lior Wolf was awarded the 2008 Sackler Career Development Chair, the Colton Excellence Fellowship for new faculty (2006-2008), the Max Shlumiuk award for 2004, and the Rothchild fellowship for 2004. His joint work with Prof.

Shashua in ECCV 2000 received the best paper award, and their work in ICCV 2001 received the Marr prize honorable mention. He was also awarded the best paper award at the post ICCV workshop on eHeritage, 2009.