How We Look Tells Us What We Do: Action Recognition Using Human Gaze

Kiwon Yun¹, Gary Ge², Dimitris Samaras¹, Gregory J. Zelinsky¹,³

¹Department of Computer Science, Stony Brook University, ²Ward Melville High School, ³Department of Psychology, Stony Brook University

Big Picture
Eye movements contain information that can be used to recognize actions in still images and enhance automatic computer vision methods.

Information in Eye Movements
• Different action classes elicit different spatio-temporal gaze patterns from viewers.
• Gaze features are derived and used to train Support Vector Machine (SVM) classifiers.
• Confusion in the gaze classifier reveals behaviorally-meaningful action groups.

Information in Pixels
• Convolutional Neural Network (CNN) features are computed for an image and are used to train SVM classifiers for each action.
• Explore relationship between gaze patterns and pixels describing actions in images.
• Show usefulness of gaze, alone or combined with computer vision, to classify images.

Goal
• Better understand through gaze how people comprehend and group actions.
• Propose novel gaze features for automatic action classification in still images.

Contribution
• Gaze features are derived and used to train Support Vector Machine (SVM) classifiers.

Datasets
PASCAL VOC 2012 Action Classes

Gaze Data
• 500 images selected from a total of 9157 images featuring:
  - 10 action classes: "walking", "running", "jumping", "riding horse", "riding bike", "phonning", "taking photo", "using computer", "reading", and "playing instrument".
  - All selected images depicted a single whole person performing an action.

Experiments & Analyses
Visualizing gaze patterns

Aggregate fixations from all subjects, with darker circles denoting earlier fixations.

Fixations clustered with a Gaussian Mixture Model.

Gaze Density Maps (FDMs) using 2D Gaussian distributions weighted by fixation duration.

Classifications results for 10 action classes
• Separate SVM classifiers were trained using gaze and CNN features.
• 2 different versions of baseline: CNN and CNN-MultiReg.
• Gaze and baseline were combined by summing weighted confidence scores.

Classifiers for individual action classes

Eye tracker

Gaze Data

Visualization

Classifiers for combination of gaze and CNN features

Eye movement data collected by [1]
- 8 subjects (3 male and 5 female)
- 3 second viewing period.
- Task: Recognize the action in an image and select it from a list of 10 actions.

Acknowledgements
This work was supported in part by NSF grants IIS-1161876, IIS-1111047, and the SubSample Project by the DIGITEO institute, France. We thank Minh Hoai for providing pre-computed CNN features.


Presented in VSS 2015. The paper version of this work can be found as the following title: “Action Classification in Still Images Using Human Eye Movements”, Vision Meets Cognition Workshop, CVPR 2015.