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

Kiwon Yun[1], Gary Ge[2], Dimitris Samaras[1], Gregory J. Zelinsky[1,3]
[1]Department of Computer Science, Stony Brook University, [2]Ward Melville High School, [3]Department of Psychology, Stony Brook University

Stony Brook University

EYE COG LAB

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

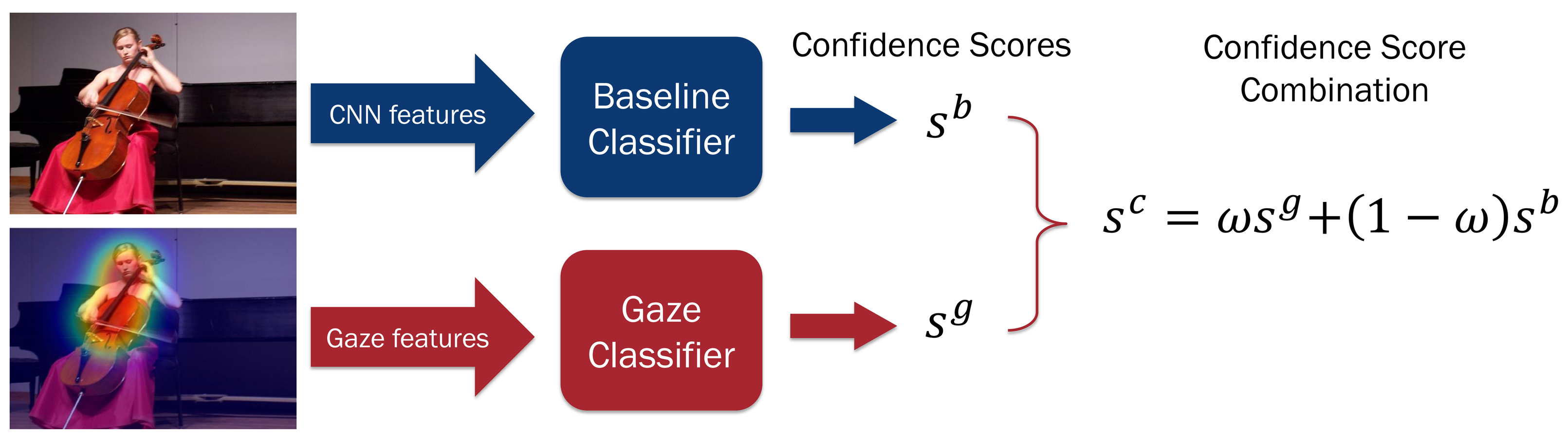
- Different action classes elicit different spatio-temporal gaze patterns from viewers.
- Gaze features are derived and used to train Support Vector Machine (SVM) classifiers.
- Confusion in the gaze classifier reveals behaviorally-meaningful action groups.

Information in Pixels

- Convolutional Neural Network (CNN) features are computed for an image and are used to train SVM classifiers for each action.

Goal

- Explore relationship between gaze patterns and pixels describing actions in images.
- Show usefulness of gaze, alone or combined with computer vision, to classify images.



Contribution

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

Datasets

PASCAL VOC 2012 Action Classes

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

Gaze Data

- 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

[1] C. S. Stefan Mathe. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In Advances in Neural Information Processing Systems, 2013.

Experiments & Analyses

Visualizing gaze patterns

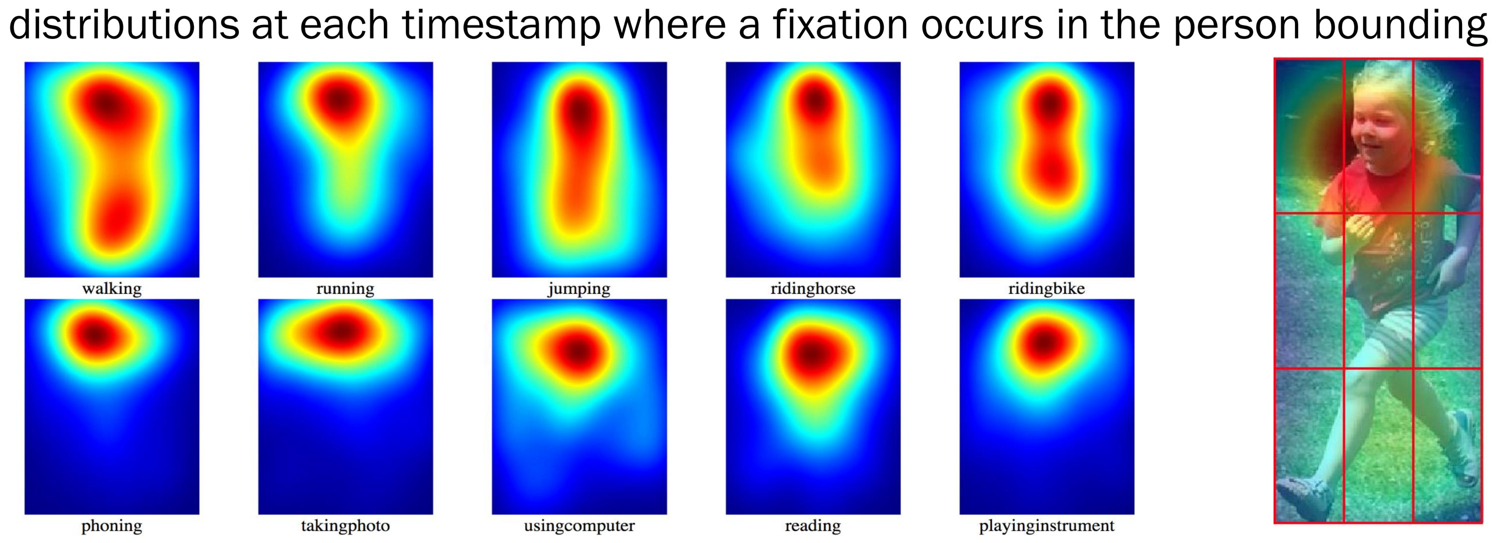
Aggregate fixations from all subjects, with darker circles denoting earlier fixations.

Fixations clustered with a Gaussian Mixture Model.

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

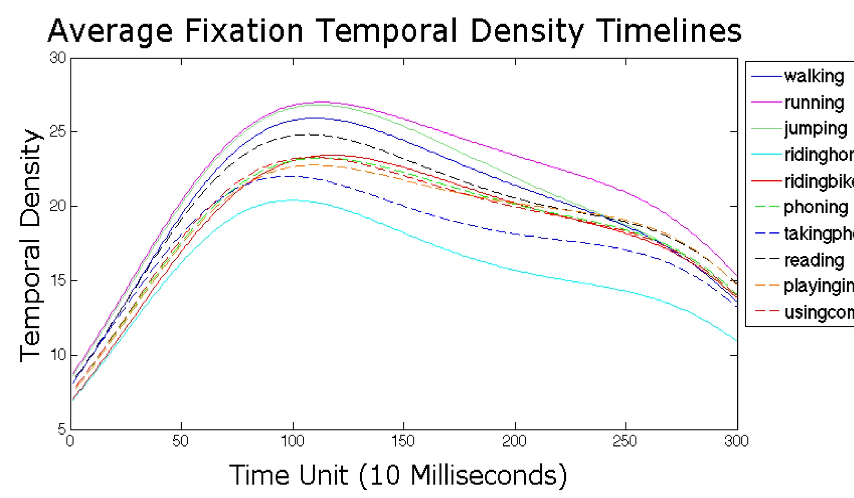
Gaze Features

- Some features used 9 grid regions, others used 3 (upper-body/lower-body/context).
- Fixation Density Maps (FDMs) generated by placing duration-weighted 2D Gaussian distribution at the location of each fixation on an image.
- Transitions between upper-body, lower body, and context segments were measured.
- Temporal Density Timelines generated by placing duration-weighted Gaussian distributions at each timestamp where a fixation occurs in the person bounding box.



Average FDMs for each action class, and the 9 segments from which features are extracted.

Gaze transitions measured between upper-body, lower-body and context segments.



Subjects looked at the person in the image within a second, then looked to context relevant objects.

Gaze Feature	Dim.
9 Region (FDM mean/max)	18
9 Region (transitions)	36
Upper/Lower/Context (fixation duration)	3
Upper/Lower/Context (transitions)	3
Temporal Density Timeline (mean/max)	12

Summary of feature types and dimensions.

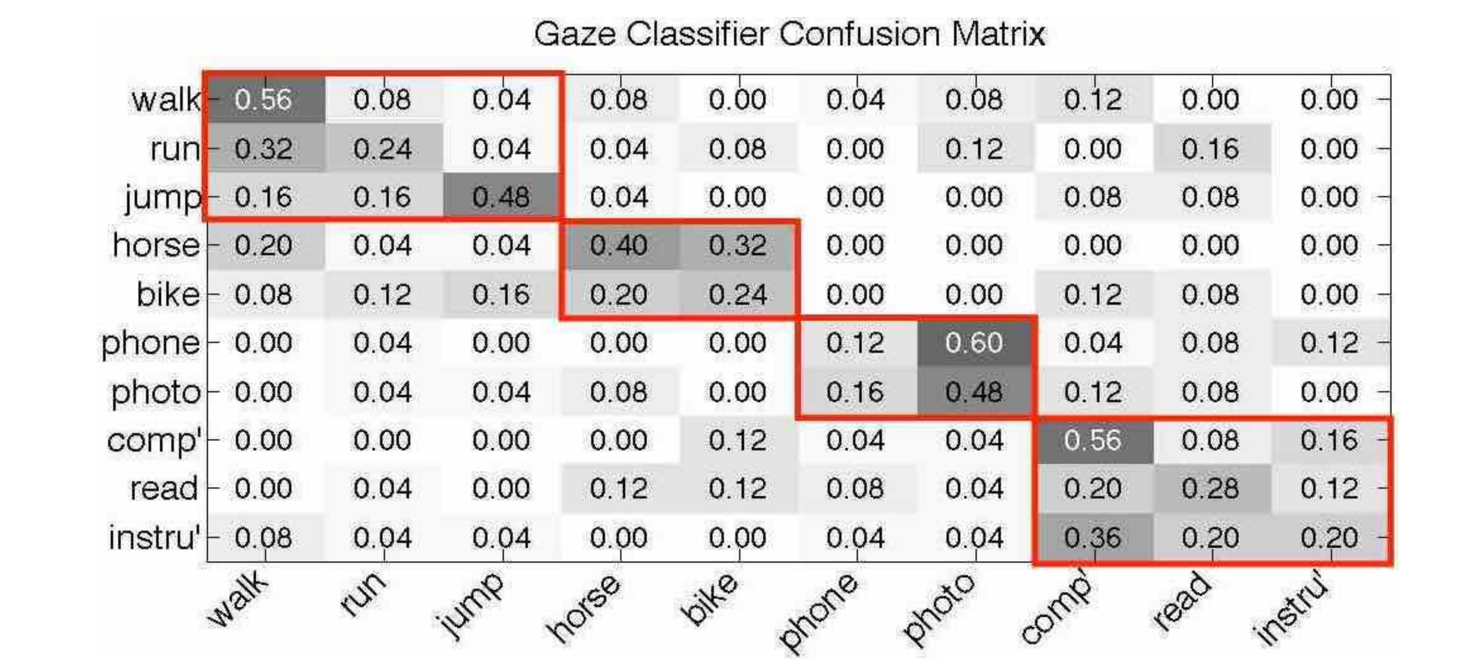
Classification results for 10 action classes

- 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.

	Gaze Features	CNN	CNN-MultiReg	Gaze + CNN	Gaze + CNN-MultiReg
walk	46.72	35.22	**58.03**	35.22	**58.03**
run	41.75	74.69	**77.70**	74.68	**77.70**
jump	41.65	74.03	**87.47**	78.59	**87.47**
horse	70.63	91.22	**98.41**	92.99	94.75
bike	34.15	**98.70**	96.63	**98.70**	96.63
phone	47.58	36.20	**49.29**	36.20	**49.29**
photo	46.24	42.53	**57.94**	42.54	**57.94**
comp'	38.74	**74.34**	72.84	**74.34**	72.84
read	35.01	59.73	58.46	**60.19**	58.46
instru'	36.08	60.95	**67.24**	60.96	**67.24**
mAP	43.86	64.76	**72.40**	65.44	72.04

Average Precisions (APs) for classification of 10 actions. Higher APs are bolded.

Gaze Classifier Confusion Matrix



The confusion matrix shows four groups of commonly-confused classes that are behaviorally meaningful. We retrained classifiers to discriminate between these groups.

Classification results for four class groups

- SVM classifiers were retrained to discriminate between four action groups:
  - walking + running + jumping
  - riding horse + riding bike
  - phoning + taking photo
  - using computer + reading + playing instrument

	Gaze Features	CNN	CNN-MultiReg	Gaze + CNN	Gaze + CNN-MultiReg
walk + run + jump	80.33	86.39	88.72	**92.29**	90.21
horse + bike	79.21	97.53	97.63	**98.99**	98.32
phone + photo	**81.64**	61.13	65.35	76.09	76.36
comp' + read + instru'	83.48	92.21	92.32	93.93	**94.10**
mAP	81.17	84.32	86.01	**90.33**	89.75

Average Precisions (APs) for classification of four action groups. A combination of gaze and CNN features performs best overall. Higher APs are bolded.

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