Action Classification in Still Images Using Human Eye Movements

Gary Ge
Ward Melville High School, East Setauket, NY 11733, USA
garyliangge@gmail.com

Kiwon Yun Dimitris Samaras Gregory J. Zelinsky
Stony Brook University, Stony Brook, NY 11794, USA
{kyun,samaras}@cs.stonybrook.edu, {gregory.zelinsky}@stonybrook.edu

Abstract

Despite recent advances in computer vision, image categorization aimed at recognizing the semantic category of an image such as scene, objects or actions remains one of the most challenging tasks in the field. However, human gaze behavior can be harnessed to recognize different classes of actions for automated image understanding. To quantify the spatio-temporal information in gaze we use segments in each image (person, upper-body, lower-body, context) and derive gaze features, which include: number of transitions between segment pairs, avg/max of fixation-density map per segment, dwell time per segment, and a measure of when fixations were made on the person versus the context. We evaluate our gaze features on a subset of images from the challenging PASCAL VOC 2012 Action Classes dataset, while visual features using a Convolutional Neural Network are obtained as a baseline. Two support vector machine classifiers are trained, one with the gaze features and the other with the visual features. Although the baseline classifier outperforms the gaze classifier for classification of 10 actions, analysis of classification results over reveals four behaviorally meaningful action groups where classes within each group are often confused by the gaze classifier. When classifiers are retrained to discriminate between these groups, the performance of the gaze classifier improves significantly relative to the baseline. Furthermore, combining gaze and the baseline outperforms both gaze alone and the baseline alone, suggesting both are contributing to the classification decision and illustrating how gaze can improve state of the art methods of automated action classification.

1. Introduction

Visual media such as images are powerful forms of communication due to the vast amount of information they can convey, but image understanding remains a tremendously complex and nuanced task that is only reliably conducted by humans. In particular, eye movement behavior is strongly connected with the way humans interpret images and has been widely studied in cognitive science [28, 8]. Recently, several researchers have started to use eye movement data in conjunction with automatic computer vision algorithms for better understanding of images and videos. Human gaze is used for many standard computer vision tasks such as image segmentation [17, 19], object detection [30, 29, 18], and face and text detection [11], but the usefulness of human gaze for action recognition in still images has not beeen studied yet. There exists a vast number of applications for gaze-enabled action classification, including image retrieval, real world action detection (as with security footage), and image annotation of large-scale datasets such as ImageNet, Flickr, etc. Mathe and Sminchisescu [21] propose prediction of gaze patterns for different behavioral tasks, but eye movement data has not been directly used for action classification. Even though some work has shown that eye movement behavior leads to improved performance for action recognition [16, 22] and action localization [20] in videos, human gaze has only been used to identify salient regions on which computer vision features are then computed. Furthermore, sophisticated spatio-temporal gaze features have not been studied for action recognition. Thus, the focus of this study is to explore the usefulness of gaze data for action recognition in still images.

Recognizing from a still image the action that humans are performing is one of the most difficult facets of image analysis, and a multitude of methods has been developed in recent years. Reserachers have used low-level visual features (e.g SIFT [14] and HOG [3]), high-level features (e.g. attributes [27, 1], body parts and pose [25, 10, 9] and human-object interaction [26, 24]) and more sophisticated visual features such as Convolutional Neural Network [2, 12]. Furthermore, the Bag-of-Words (BoW) [4] repre-

sentation combined with Spatial Pyramid Matching (SPM) [13] or Regularized Max Pooling [9] has been often used.

The PASCAL Visual Objects Classes (VOC) 2012 Actions image set is one of the most difficult publicly available datasets for action classification. Ten different action classes, plus a trivial 'other' class, are identified: 'walking', 'running', 'jumping', 'ridinghorse', 'ridingbike', 'phoning', 'takingphoto', 'usingcomputer', 'reading', and 'playinginstrument'. We analyze eye movement data collected from Mathe and Sminchisescu [21], and propose several novel gaze features for action classification. In order to quantify behavioral patterns under simplified conditions, we choose a subset of images containing exactly one whole human performing an action. We automatically split the annotations as segments (e.g. person, upper-body, lower-body, context) and derived gaze features, which includes: number of transitions between segment pairs, avg/max of fixation-density map per segment, dwell time per segment, and a measure of when fixations were made on the person versus the context.

Two Support Vector Machine (SVM) classifiers are trained, one using visual features and one using gaze features, and the confidence scores of the classifiers are combined in a novel combination method. Although average precision across the ten action categories was poor, the gaze classifier revealed four distinct behaviorally-meaningful subgroups where actions within each subgroup were highly confusable. Retraining the classifiers to discriminate between these four subgroups resulted in significantly improved performance for the gaze classifier. Moreover, the gaze+vision classifier outperformed both the gaze-alone and vision-alone classifiers, suggesting that gaze-features and vision-features are each contributing to the classification decision.

From a cognitive psychology perspective, we attempt to analyze patterns in gaze across various actions and look for inter-class differences as well as intra-class consistency through spatial, temporal, and durational features. Our results have implications for both behavioral and computer vision; gaze patterns can reveal how people group similar actions, which in turn can improve automated action recognition.

The paper is organized as follows: Section 2 provides a detailed description of the dataset. In Section 3, we propose several novel gaze features and describe the classification methods. Section 4 shows the experimental results, and the reasons for and meanings behind these results are discussed in Section 5. Lastly, Section 6 concludes the paper.

2. Datasets

Images: The challenging PASCAL VOC 2012 Action Classes dataset is divided into a training and validation set of 4588 images and a test set of 4569 images for a total of 9157 images [5]. However, these images contain a large amount of intra-class variation regarding the number of humans in an image, the number of actions being performed in an image, and the fraction of the human found in an image (such as a whole human, only a torso, only a head, or almost none at all). For gaze data to be meaningful and comparable across images, 50 images per class (for a total of 500) were selected to depict exactly one whole person performing an action. Some iterations of the action classification challenge allow for annotations such as human bounding boxes on the bodies of the humans [5]. The baseline computer vision algorithm utilizes human bounding box annotations, so for the sake of consistency, gaze features were computed using the same annotations.

Eye Movement Data: Fixation data [21] were collected with an SMI iView X HiSpeed 1250 tower-mounted eye tracker over the entire PASCAL VOC 2012 Action Classes dataset from 8 subjects who were asked to recognize the actions in the images and indicate them from the labels provided by the PASCAL VOC dataset. Subjects were given 3 seconds to freely view an image, during which x- and y-coordinate gaze position was recorded. The first of these fixations is also discarded because subjects started each trial by fixating a cross corresponding to the center of each image. Figure 1 shows an example of human eye movement data in the dataset.

3. Action Classification using Gaze Features

3.1. Analysis of Human Gaze

To discover gaze patterns, we explored different methods of visualizing the fixation data. We first visualize gaze data by plotting fixations for each subject and for each image in the dataset, with saccades indicated by lines and order indicated by color temperature. Aggregated fixation visualizations are also generated by plotting fixations for all subjects on the same image. Spatial agreement between subjects is indicated by more tightly-grouped fixations, while temporal agreement is indicated by similar patterns of saccade lines (Figure 1b). Fixations from all subjects were also clustered using a Gaussian Mixture Model (GMM) [23]. Denser fixation clusters suggest areas of interest with high subject agreement, whereas sparser fixation clusters suggest idiosyncratic differences in viewing behavior perhaps related to establishing scene context (Figure 1c). Lastly, we generate fixation density maps (FDMs) computed using a two-dimensional Gaussian distribution, with a sigma corresponding to one degree of visual angle, centered on each fixation and weighted by fixation duration. (Figure 1d).

By exploring the visualizations, we made several observations that could be useful for classification. First, the number of fixations and the dwell time over the upper-body and lower-body are significantly different between classes

Figure 1: Various methods of visualization are used to elucidate gaze patterns. Examples of original images from the dataset are shown in a. Aggregated fixations are illustrated in b, with different subjects distinguished by color. Earlier fixations are drawn with darker circles, while later fixations are drawn with lighter circles. Results of fixation clustering with GMM are shown in c with different colors representing different clusters. FDMs weighted by fixation duration are illustrated in d.

(Figure 3), as is the number of fixations over the context (Figure 1c). Second, the number of transitions is also different. Moreover, the FDMs are clearly different among action classes (see Figure 2, therefore, the FDM is a possible gaze feature for the action classification. The temporal orders of fixations are important information we must convey in a feature as well.

3.2. Gaze Features

Given these observations, we propose several novel gaze features for action classification. A heuristic approach was taken to divide an image into useful segments for gaze analysis. Upper- and lower-body segments of the person in each image were identified by taking the ratio of the width and height of the human bounding box and splitting the bounding box either horizontally or vertically for those with greater heights or widths, respectively. In images with horizontally-split bounding boxes, the upper segment is assumed to be the upper body and the lower segment is assumed to be the lower body. However, images with vertically-split bounding boxes lack an intuitive method of identifying the upper- and lower-body segments; instead a head detector [15, 6] is run in each segment and the higher-scoring segment is considered the upper body. The resulting three segments (upper-body, lower-body and context) are used to obtain the number of gaze transitions between each pair of segments. The total duration of fixations in each segment is also computed (Figure 3).

Similarly, the human bounding box is divided into nine equal segments and the number of transitions between each pair of segments is calculated for a total of 36 gaze features. The sparseness of these features is compensated by the higher number of features generated by this method. FDMs are also generated for the image and the mean and max values of the bounded portion of the FDM are taken for each subregion of the human bounding box (Figure 5).

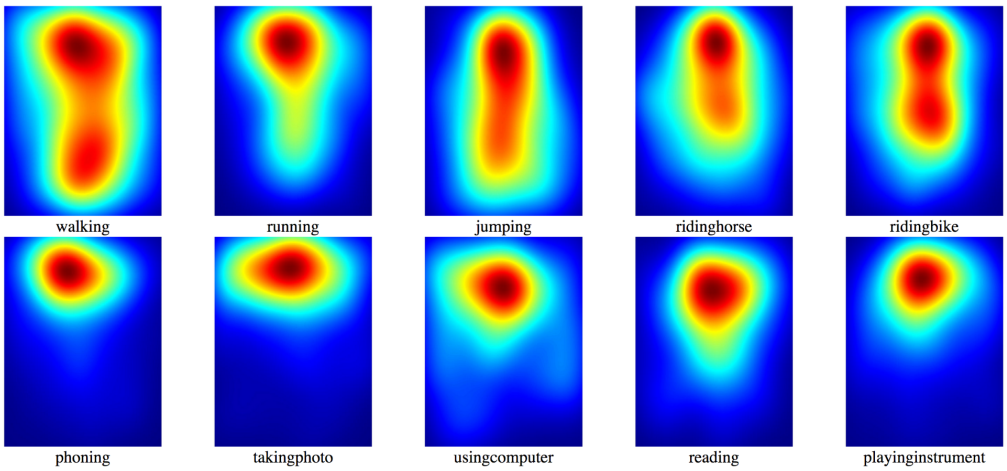

Figure 2: Average FDMs generated from regions of duration-weighted FDMs bounded by human bounding boxes. Transitions between upper and lower body are shown in the 'walking', 'running', 'jumping', 'ridinghorse' and 'ridingbike' classes.

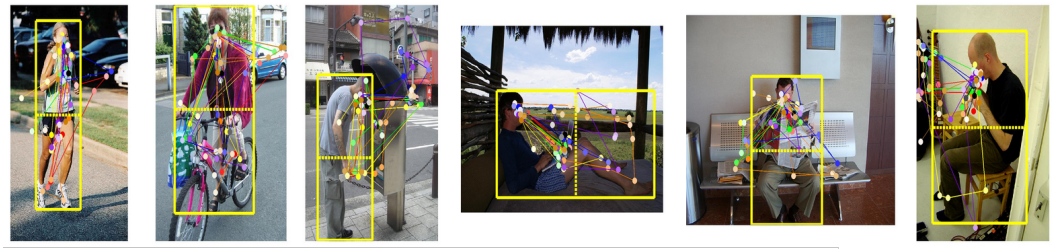

Figure 3: Gaze transitions between the upper body segment, the lower body segment and the context segment are counted for each subject, yielding 3 gaze features.

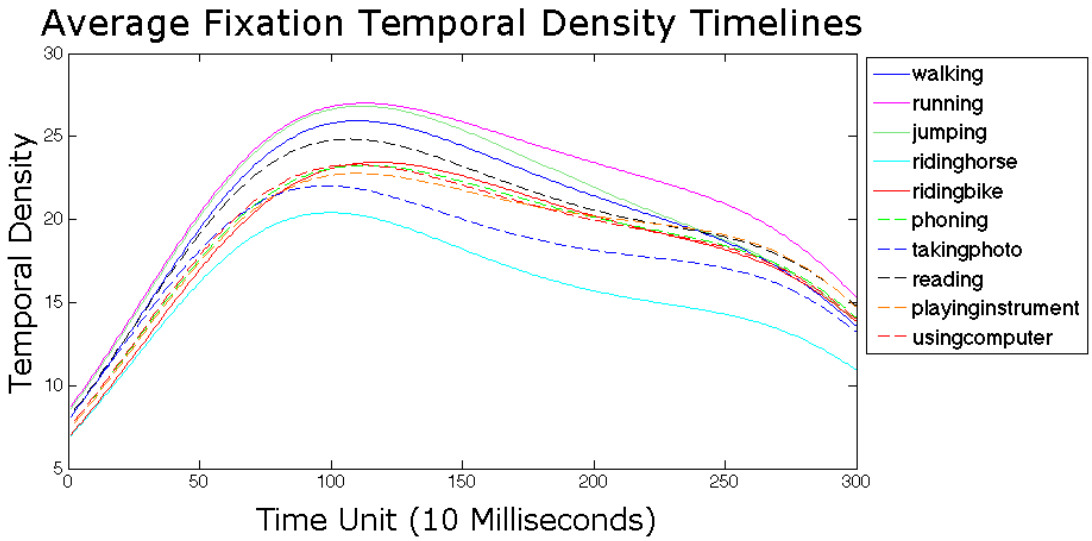

Figure 4: Average fixation temporal density timelines from all images in the dataset for each class are illustrated. The mean and maximum values in each segment are taken as features. Subjects looked at the person in the image within a second, resulting in the peak observed at that time. The following decline in temporal density likely reflects later looks to context-relevant objects. Although temporal density differs between classes, all classes showed a generally similar trend. Figure best viewed in color.

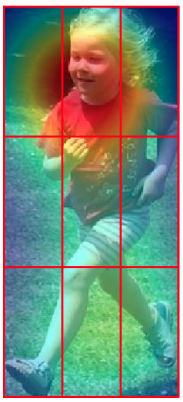

Figure 5: Fixation density maps were split into nine equal segments, from which the mean and maximum values of each segment and the number of transitions between each pair of segments are used as features.

Gaze Feature	Dimension
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

Table 1: We create several gaze features using different sets of spatial segments (9 Region and Upper/Lower/Context) as well as temporal information (Temporal Density Timeline). Dimensions of each feature type are given.

Lastly, fixation temporal density is quantified in a manner similar to spatial density by associating a vector of 300 values to each image to represent duration of the viewing period in units of 10 milliseconds. A Gaussian distribution is placed at each timestamp where a fixation within the human bounding box occurs, and the vector is divided into equal segments from which the mean and maximum values are taken as features (Figure 4). Six segments were chosen to balance sparseness and resolution.

3.3. Action Classification

Support Vector Machine (SVM) classifiers are trained for action recognition in still images. For the training set we randomly select 25 images from each class for a total of 250 images; the remaining images constitute the testing set. In the process of training, a random subset of 50 images is selected from the training set as a validation set to tune various parameters with 5-fold cross validation. For the baseline, SVM is trained with a linear kernel, since the CNN features with fixed subregions were demonstrated to exhibit no significant difference between classification with a linear kernel and with an RBF kernel [9]. A polynomial kernel was used with the gaze features, since this kernel was demonstrated to perform best overall. All features were normalized before training SVMs. We conduct 1-vs-all binary classification for each action and measure performance using Mean Average Precision (mAP), which is a standard measurement used for many visual recognition challenges [8]. To obtain mAP, we first compute the Average Precision (AP) for each of the action classes, and the mAP is the average of the obtained APs. A confusion matrix was generated for the gaze classifier to determine which actions were commonly confused. Frequently confused actions were grouped and SVMs were retrained to account for the confusion.

3.4. Baseline and Feature Combination

For baseline comparison we derive purely visual features using a Convolutional Neural Network (CNN). CNN features have a substantially more sophisticated structure than standard representations using low-level features, and have shown excellent performance for image classification [2, 12]. To extract CNN features, we use the publicly available Caffe implementation [2] of the CNN architecture described by Krizhevsky et al [12]. There exist two different baselines. For the first baseline, we extract a 4096 dimensional CNN feature vector from the human bounding box. This feature vector is referred to as CNN_M_4096 in [2]. For the second baseline, we extract CNN features for 100 fixed subregions as in [9]. An image is divided into a grid of 16 blocks (4x4) and analyzed in terms of 100 rectangular subregions formed by contiguous chunks of blocks. This method is called MultiReg in [9]. MultiReg is similar to SPM, but it is based on rigid geometric correspondence and features from a different set of subregions. The second baseline is able to convey a spatial relationship between body parts of the person.

The combination method is followed by Equation 1.

$$\mathbf{s^c} = \omega \mathbf{s^g} + (1 - \omega)\mathbf{s^b} \tag{1}$$

where $\mathbf{s^g}$ and $\mathbf{s^b}$ are SVM confidence scores for the gaze and baseline classifiers, respectively, while $\mathbf{s^c}$ is the SVM confidence score of the combination method. We optimize ω with 5-fold cross validation on training images and use $\mathbf{s^c}$ to compute average precision.

4. Experimental Results

Classifiers were trained using the CNN features computed for 100 fixed subregions as well as using only the CNN features corresponding to the entire human bounding box (HBB). The combination method was implemented using both CNN feature classifiers as the 'baseline' classifier. The computed APs for the gaze classifier were found to be

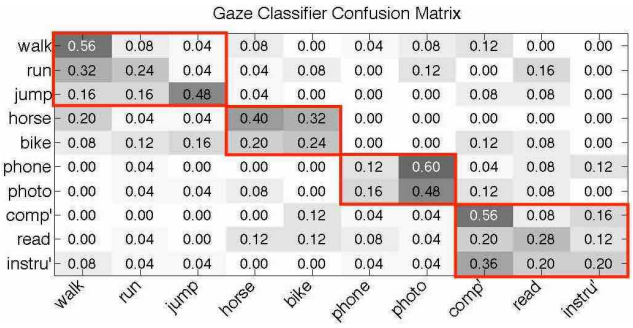

Figure 6: The confusion matrix for the gaze-alone classifier suggests that certain similar actions tend to elicit similar gaze patterns. Classification results improve when these actions are grouped as illustrated above.

comparable to those of the CNN classifiers for the 'walking', 'phoning' and 'takingphoto' classes (Table 2). APs for confidence vector combination were very close to those of the CNN feature classifiers and the combination using CNN human bounding box features actually performed best in the 'reading' class, but the CNN classifier using 100 subregions performed best overall.

Figure 6 shows the confusion matrix of the result. The actions were divided into 4 groups of commonly confused classes: 1) 'walking', 'running', and 'jumping'; 2) 'ridinghorse' and 'ridingbike'; 3) 'phoning' and 'takingphoto'; and 4) 'reading', 'usingcomputer', and 'playinginstrument'.

4.1. Classification Results with Subgroups

The classifiers were retrained with new ground truth labels associated with the aforementioned commonly confused classes. The performance of the gaze classifier was comparable to that of the CNN classifiers in most groups and actually outperformed them in the 'phoning + takingphoto' group. Moreover, the combination using CNN features over 100 subregions performed best in the 'usingcomputer + reading + playinginstrument' group, while the combination using CNN human bounding box features performed best in the remaining two class groups and performed best overall, demonstrating that the best classification results arise from a combination of gaze and CNN features at the classifier level (Table 3).

5. Discussion

The gaze classifier yielded results that were comparable to those of the CNN feature classifiers for single action classification, so gaze-based classifiers have the potential to serve as an alternative action classification method. When frequently confused classes are grouped together and classifiers are retrained, the performance of the gaze classifier improves dramatically and the method of combining confidence outputs with optimized weights significantly outperforms the CNN classifiers, indicating that gaze can be used as an effective supplement to existing visual features and can give unique insight into the task of action classification.

5.1. Gaze Classifier Confusion

Figure 6 indicates some degree of confusion between classes using the gaze-alone classifier. The most frequent instances of confusion occur among the actions grouped into the aforementioned four classes. The 'walking', 'running', and 'jumping' classes tend to be mistaken for each other since the poses of people performing these actions are very similar. The human process of discriminating between these classes also yields similar gaze results. In particular, subjects looked at faces, which were naturally interesting areas of the image, and the legs, which contained information about which of the three actions the person was performing. Thus, for all three classes there were notable transition patterns from the upper- and lower-body segments of the human bounding box.

The natural similarities between the 'ridinghorse' and 'ridingbike' classes also accounted for similar gaze patterns, as subjects looked below the people in images to identify the horses or bikes they were riding. In certain images where the horse or bike rider was facing sideways, subjects also looked at the hands of humans riding bikes in the same manner as they looked at the heads of horses, both of which were located in front of the human and gave information about the action being performed in the image. Without any other information about the image, these similar patterns contributed to the confusion between these two classes. Gaze transitions between heads and legs of the 'walking', 'running', and 'jumping' classes were also similar to those between human heads and the horses or bikes in the 'ridinghorse' and 'ridingbike' classes, respectively, so confusion between all five classes was also observed.

The 'phoning' and 'takingphoto' classes, both involving small devices commonly held near the head, were frequently confused since fixations were clustered around the head for most images. However, since fixations on phones/cameras were largely indistinguishable from fixations on the head, gaze patterns for the 'phoning' and 'takingphoto' classes were perceived to exhibit much greater frequencies of head fixations than those for any other classes. Thus, using gaze for classification of these two actions grouped together yielded excellent results that outperformed the baseline significantly (Table 3).

Since the 'usingcomputer', 'reading' and 'playinginstrument' classes all involved interaction with an object that was generally held away from the body, they prompted gaze transitions between the head and hands. Since there tended to be greater deformation in these classes than in any oth-

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

Table 2: AP is computed for each classifier and for each action. Gaze is comparable with CNN in some classes, and combination methods yield similar results as CNN, suggesting the baseline confidence vectors are contributing most to the decision. Highest AP values for each class are bolded, as is the highest mAP.

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

Table 3: AP values are computed for each classifier and for each group of commonly-confused classes. Gaze performs best for the 'phoning + takingphoto' group, while some combination of gaze and CNN performs best for all other groups. The highest AP value for each group of classes is bolded, as is the highest mAP.

ers, intra-class variation in gaze patterns resulted in difficult classification and high confusion between these three classes and, to a lesser extent, with other classes. In particular, confusion between these three classes and the 'ridingbike' class is somewhat higher since gaze transitions between the head and hands occur to find handlebars of bikes.

6. Conclusions and Future Work

We proposed several novel gaze features for action classification on still images. The efficacy of gaze as a basis for action classification was illustrated by its comparable results to a state of the art computer vision algorithm and the combination method demonstrated potential to outperform both the gaze-alone and vision-alone classifiers, suggesting that gaze-features and vision-features are each contributing to the classification decision. These results have implications for both behavioral and computer vision; gaze patterns can reveal how people group similar actions, which in turn can improve automated action recognition. With the proliferation of laptop webcams and front-facing cell phone cameras, everyday devices have the potential to record eye movement information of everyday users.

Gaze can also be used to deduce intentions of humans in the image being viewed, which is very relevant for image classification. For instance, eye movements may correlate with the direction that humans are looking and indicate their focus or may point out the locations of hands and reveal salient regions related to the task being performed. These higher-level features would benefit from more specific segments such as hands and faces, which could be found with additional annotations or effective hand and face detectors.

However, more complex methods of combining gaze with the state-of-the-art classification algorithms may offer the most promising opportunities for improving classification performance. Hoai [9] propose the Regularized Max Pooling (RMP) model, which allows for substantial deformation for geometrically analogous images. It may be possible to combine spatio-temporal information from fixations with this method for better performance. Multiple Kernel Learning (MKL) involves the use of a combination of different kernel types instead of a single, best-performing kernel [7]. A hierarchical approach may use a combination method to classify groups and visual features to classify within them. In future work we intend to use these approaches for automated action recognition.

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 precomputed CNN features.

References

[1] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV 2011*, pages 1543–1550. IEEE, 2011. 1

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC 2014*, 2014. 1, 5

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR 2005*, volume 1, pages 886–893. IEEE, 2005. 1

[4] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC 2010*, 2010. 1

[5] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2012, 2012. 2

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 3

[7] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011. 7

[8] J. M. Henderson. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504, 2003. 1, 5

[9] M. Hoai. Regularized max pooling for image categorization. In *Proceedings of British Machine Vision Conference*, 2014. 1, 2, 5, 7

[10] M. Hoai, L. Ladicky, and A. Zisserman. Action recognition from weak alignment of body parts. In *Proceedings of British Machine Vision Conference*, 2014. 1

[11] S. Karthikeyan, V. Jagadeesh, R. Shenoy, M. Ecksteinz, and B. Manjunath. From where and how to what we see. In *ICCV 2013*, pages 625–632. IEEE, 2013. 1

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 1097–1105. Curran Associates, Inc., 2012. 1, 5

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR 2006*, volume 2, pages 2169–2178. IEEE, 2006. 2

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1

[15] M. J. Marin-Jimenez, A. Zisserman, M. Eichner, and V. Ferrari. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296, 2014. 3

[16] S. Mathe and C. Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *ECCV 2012*, pages 842–856. Springer, 2012. 1

[17] A. Mishra, Y. Aloimonos, and C. L. Fah. Active segmentation with fixation. In *ICCV 2009*, pages 468–475. IEEE, 2009. 1

[18] D. P. Papadopoulos, A. D. Clarke, F. Keller, and V. Ferrari. Training object class detectors from eye tracking data. In *ECCV 2014*, pages 361–376. 2014. 1

[19] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. An eye fixation database for saliency detection in images. In *ECCV 2010*, pages 30–43. 2010. 1

[20] N. Shapovalova, M. Raptis, L. Sigal, and G. Mori. Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. In *Advances in Neural Information Processing Systems*, pages 2409–2417, 2013. 1

[21] 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. 1, 2

[22] E. Vig, M. Dorr, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *ECCV 2012*, pages 84–97. 2012. 1

[23] G. Xuan, W. Zhang, and P. Chai. Em algorithms of gaussian mixture model and hidden markov model. In *Proceedings of International Conference on Image Processing*, volume 1, pages 145–148. IEEE, 2001. 2

[24] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR 2010*, pages 9–16. IEEE, 2010. 1

[25] B. Yao and L. Fei-Fei. Action recognition with exemplar based 2.5 d graph matching. In *ECCV 2012*, pages 173–186. 2012. 1

[26] B. Yao and L. Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1691–1703, 2012. 1

[27] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV 2011*, pages 1331–1338. IEEE, 2011. 1

[28] A. L. Yarbus, B. Haigh, and L. A. Rigss. *Eye movements and vision*, volume 2. Plenum press New York, 1967. 1

[29] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Exploring the role of gaze behavior and object detection in scene understanding. *Frontiers in psychology*, 4, 2013. 1

[30] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Studying relationships between human gaze, description, and computer vision. In *CVPR 2013*, pages 739–746. IEEE, 2013. 1