Determining child orientation from overhead video: a multiple kernel learning approach

Marie D. Manner, Ming Jiang,
Qi Zhao, Maria Gini
College of Science and Engineering
University of Minnesota
Email: manner@cs.umn.edu, mjiang, qzhao, gini@umn.edu

Jed Elison
Institute of Child Development
University of Minnesota
Email: jtelison@umn.edu

***Abstract*—Our goal is to automatically detect which direction a child is facing based on a single, simple overhead picture, and track that direction across time. Engaging in joint attention, which is the shared focus of two individuals on some object of interest, is a strong cue of typically developing children, and the lack thereof can be an indicator of autism spectrum disorder or other pervasive developmental disorder. Therefore, the goal of many psychology experiments with children is to determine when, for how long, and towards what the child looks after some bid for attention or reaction. While much research looks for the orientation of faces based on frontal or profile pictures, or non-morphable, larger objects like cars, fewer studies work in the setting of minimally-invasive overhead person gaze or orientation detection. To automatically detect the child's orientation during a human-robot interaction experiment, we mount a camera on the ceiling of a child development laboratory and analyze the video footage. We use multiple kernel learning on eight potential orientation directions to determine a child's orientation during the video recorded interaction. We also contribute the labelled dataset we used on this challenging problem.**

I. Introduction

Our work is part of a larger experimental project in which we analyse the reaction of a child while s/he plays games with a robot. The overarching goal is to assist in the detection of abnormal development by leveraging the interest of children with autism in robots, as current literature [1] demonstrates. As such, the children we work with may be on the autism spectrum or have another pervasive developmental disorder, and we want the interaction to be as natural as possible and allow the child to be as mobile as desired. During long and occasionally mentally strenuous, structured assessments, the child participant may be calm and responsive, playful, or even destructive. Thus, instrumenting the child with technology or fiducial markers such as sensors or a hat may be infeasible; the child may not want to wear new things, may incidentally destroy any sensors within reach, or may remove the marker half way through the experiment. Any robust tool for on- or off-line interaction analysis must therefore be as non-invasive as possible. Similarly, during naturalistic play, we can not say for certain where the child will go in the room during assessments, and we want to encourage normal behaviors, such as seeking or avoiding movements as when moving towards or away from a caregiver or interesting toy. Given these challenges, analyzing overhead video recordings of interactions is the easiest and cheapest way to capture the attention and focus of our participants.

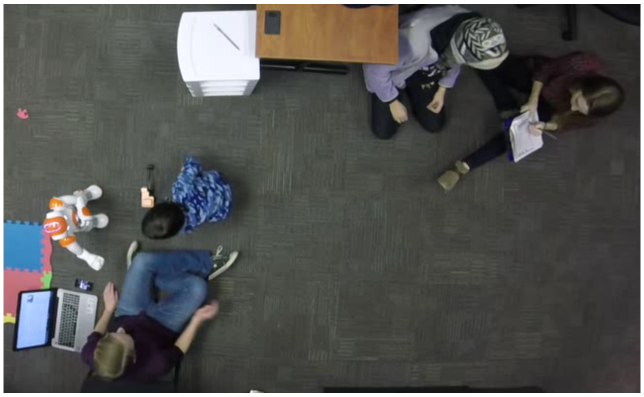
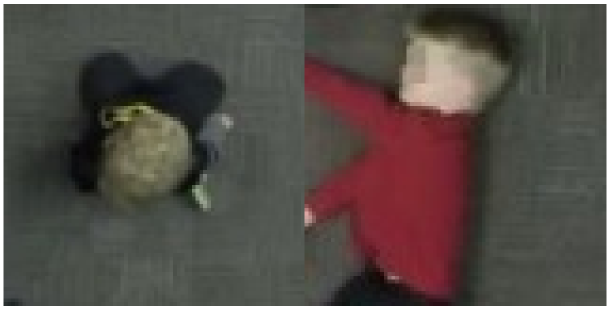


Fig. 1: Sample overhead view of the human-robot interaction experiment.

Child attention and orientation is significant in psychology for several reasons. Symptoms of some pervasive developmental disorders, such as autism spectrum disorder (ASD), include differences in personal space between individuals and objects or people [2], eye contact, physical contact, and a longer delay or non-response when called by name, among other differences [3]. These behaviors are thus considered early markers for autism, and identifying autism early in life allows for earlier treatment and far better outcomes for those with an ASD [4].

As stated earlier, we want interactions to be as natural and our sensors to be as non-invasive as possible. Close up video recordings of such development assessments, as in [5], depend on the camera sitting several feet away from the participant with the participant seated in a chair, which can be infeasible when working with active toddlers (the 2-3 year old age group), our target age for interaction analysis. Cameras, which are cheap, commercially available, and simple, may be placed anywhere out of reach of participants. The only place certain to capture the child's location, orientation, and therefore attention at any point in time is the view from the ceiling in the center of the room. Attaching a small enough camera means using simple hardware like a GoPro, which results in a variety of object perspectives in the captured data as well as distortion, but still allows for identifying individual orientation. Fig. 1 shows this vantage point, in which the child (in camouflage-patterned blue) is facing just to the robot's right (possibly at its hand), an experimenter is facing the child, and the child's parent (in a gray and white patterned hat) and another experimenter are gazing at a small stack of papers. The challenge here is to automatically detect from this video frame where each person is looking, which can be inferred from their orientations.

The view of the child, or any individual of interest, may be from directly overhead if the person is directly underneath the camera, or from an angle, if the person is anywhere else in the room or is laying on the floor, which happens with young participants. In the former case, the child's face will be occluded, and the view shows just the hair on the top of the child's head as well as the shoulders. In the latter case, the child's face may be completely occluded if she faces away from the camera, or completely visible if she faces towards the camera, or visible in profile or otherwise only partially visible, with some amount of upper body likely visible in the frame. Fig. 2 shows example images from the overhead view. The main challenge, then, is to ensure our detector correctly identifies the orientation from different vantages, when the features of the object will change.



(a) Top view (b) Partial side view

Fig. 2: Sample images of study participants. Face blurred for privacy.

Our work differentiates itself from related works on two aspects: first, we aim to estimate face orientation from video footage captured by a single overhead camera, where faces are mostly hidden. Second, the subjects are children, and their body poses have a larger variance than people in a standing position. Therefore, we have the least useful features to capture the face orientation. Compared with [6] that only detects side-to-side head turns between two opposite directions, with errors limited within ± 30 degrees, our model classifies eight directions, which is significantly more challenging.

To solve these challenges and determine where the child is facing at any point in time, we use multiple kernel learning (MKL) on the histogram of oriented gradients (HOG) of images of participants facing different directions. Our first contribution is a new approach to detecting person orientation from a single perspective overhead video recording with the use of MKL, and our second contribution is a dataset of participants with several thousand labelled samples.

In this paper, we start with related work in Section II. Section III contains our methods, including the larger problem context, laboratory set-up, and our new dataset. Section IV shows our results, and we discuss those results as well as future work in Section V.

II. Related Work

There are two background areas we draw from: automated behavior assessments through video analysis, and person location and orientation tracking. Literature in these areas shows us that more data, such as multiple camera angles or depth information, results in the best person tracking. Gaze orientation is most accurately performed with high quality pictures of faces or eyes and pupils.

A. Video based autism behaviors assessment

Automatically detecting atypical, pervasive developmental disorders, such as autism spectrum disorder, is a current research area in computer vision, and much work uses as much data as possible. For example, Hashemi et al. [5] analysed non-intrusive camera footage using a GoPro placed on a table, two to four feet from a clinician-child pair in which the clinician was testing the child with a disengagement of attention task and a visual tracking task. The authors went even further in [7], in which they analysed interest sharing and atypical motor behavior. The disengagement, visual tracking, and sharing interest analyses were done by estimating head motions from specific facial features, and the motor behavior analysis concentrated on arm asymmetry.

Fasching et al. [8] automatically coded activities of people with obsessive-compulsive disorders from overhead video footage in a structured lab, tracking how many times participants touched various objects. These objects, which include fixed environment features such as faucets, handles, and soap dispensers, may be assumed to be in the same place even after participant manipulation and between participants. In our laboratory, however, we work with much younger children and cannot be sure how the room changes between assessments.

Mead et al. [9] also investigated proxemics, placing a participant and researcher in discussion about a static humanoid robot. Using a video camera and depth data, they studied body pose during the experiment, training Hidden Markov Models on sensory experiences, such as voice loudness and a variety of distances to other people and environment objects. Using this multitude of features, including participant pose, the authors correctly annotated initiation and termination of conversation.

B. Person and orientation detection

Our work differs from other research in object and orientation detection by restricting ourselves to a single overhead camera; previous person orientation detection relies on more information. This data generally comes from additional cameras, which allows stereo reconstruction, or from additional sensors, e.g. the commercially available depth sensor Kinect.

Much work estimates gaze orientation by tracking eyes or faces, which requires varying levels of cooperation and correspondingly results in different levels of accuracy. Highly accurate, commercially available eye tracking systems can require calibration for each participant and thus requires thorough cooperation in subjects, such as the TOBII system [10], and also depends on the person being directly in front of the sensors. Other cooperation-free systems estimated gaze by first detecting facial features, such as work extending the Active Appearance Models [11] and Constrained Local Models [12]. Both methods use facial features, meaning any video footage must show a lot of the subject's face.

Multiple camera systems work well for tracking people and reconstructing the environment, but we cannot depend on multiple angles of our participants. Sivalingam studied a similar environment in [13], using multiple cameras and depth sensors to track children and adults in a classroom setting. This work is concerned with analysing the motions of children and tracking movements and patterns across multiple sessions, whereas the children in our assessments will not repeat the assessments and give longitudinal data. Bidwell used an overhead camera on a child in a seated, known location to track gaze orientation from zero to 180 degrees left or right in [6], but first found the orientation from another camera facing the participant and was able to keep the child directly under the overhead camera.

III. Method

A. Experimental paradigm

The overarching goal of the robot interaction study with toddlers (the age group of roughly 2 – 3 years old) is to identify children at high risk for ASD. We collect multiple data sets from each participant, which include parent questionnaires, established development assessments (e.g. the Mullen Scales of Early Learning that quantify skills such as expressive language, receptive language, and visual reception), human-robot interaction (HRI) experiments, and eye tracking data. We will ultimately have overhead video footage from 60 participants (experiments are still ongoing), aged two to four years old, and these videos range from roughly nine to 15 minutes long (depending on the child's willingness or ability to continue interacting with the robot). The original video is slightly distorted, thus we first perform an undistortion on each video and use the newly undistorted video to perform later analysis on. In each video, we use separate person tracking software to track the location of all actors in the scene, usually the child, the robot, one or both experimenters, and one or two caregivers.

During the HRI experiments, we introduce a child to a new friend Robbie the Robot (a NAO from Aldebaran Robotics). Robbie plays different games such as I Spy (a looking game that encourages the child to find objects in the room), Simon Says (a behavior imitation game that encourages the child to copy motions possible with gross motor skills like clapping and waving), and several dances. The set of games is in the same order for every child. The experimenter controlling the robot imitates some of the robot's movements, encourages the child to do the same, plays along during some of the looking games, and encourages the child to do the same.

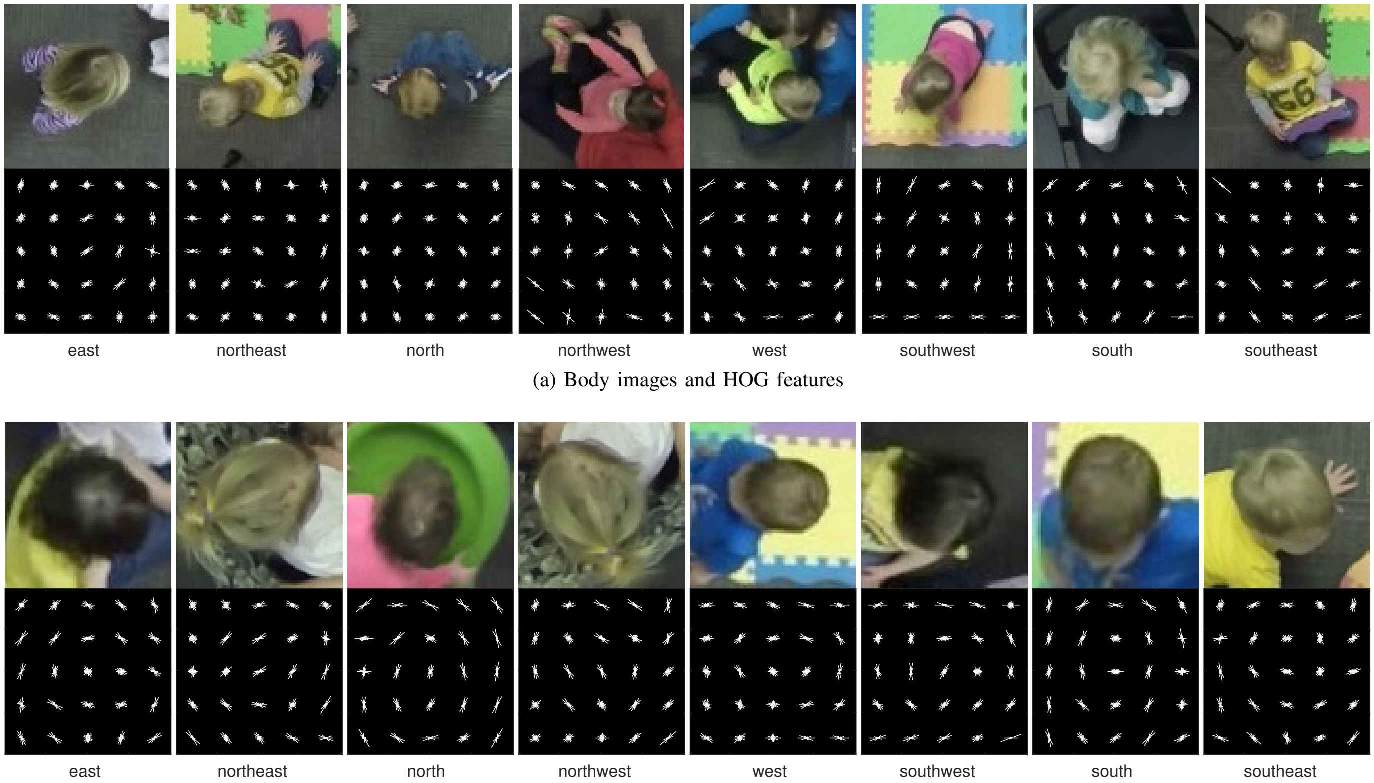
The interaction is recorded from up to four perspectives, which include from up to two sides of the room, from the robot's perspective, and from a GoPro mounted on the ceiling. The GoPro records video at resolution 1280 x 960 pixels at 30 fps. The GoPro is the only camera that is always located in the same place, and it is the only view from which we are able to see all participants in the room (which include the child, two researchers, at least one caregiver, and the robot). This work on orientation estimation is part of creating an automated response tracking method to relate the actions of the robot with reactions from the child participant. Fig. 1 shows part of a frame from an overhead view; the child participant is facing the robot, close to the researcher, and the child's parent is in discussion with another researcher.

One prong of the video footage analyses we perform is identifying when and for how long the child gazes at the robot or the experimenter. Using the overhead footage, we can track where the child faces and thus identify when the robot is an object of joint attention between the child and experimenter. We know where the child is located after processing the videos with our person tracking software, and now given the child's location in every frame, we must determine which direction the child is facing. To address this challenge, we generate a new dataset for training and testing.

B. Dataset

From the overhead video footage collection (in progress) referenced above, we take frames from ten videos using the child's location to cut out a frame of the child's full body, and, manually, cropping the child's head in a closer frame. Every image is up- or down- sampled to size 80 by 80 pixels; samples of body and head images are shown in Fig. 3. Each image was manually assigned an orientation: cardinal directions of north (straight ahead), east (to the right), south (towards the bottom of the photo) west (to the left), and in between each of those directions, the ordinal directions of northeast, southeast, southwest, and northwest.

To further augment our dataset, each photo facing north, east, south, or west was rotated in the other three directions and used as the corresponding direction. Similarly, each photo facing northeast, southeast, southwest, and northwest was rotated in the other three directions and used for that corresponding direction. Each original frame, therefore, may appear in the training dataset up to four times as different rotations. Our dataset contains over 300 samples of both ordinal and cardinal directions of overhead head photos, and over 200 samples of ordinal directions and 200 samples of cardinal directions of bodies. This data of 2400+ head samples and 1600 body samples forms our training set. We sampled another two participant videos and labelled both body and head directions to make over 50 samples of every direction; these data were not augmented by any additional rotations, so each of these 400 plus images are unique. This data forms our



(a) Body images and HOG features

(b) Head images and HOG features

Fig. 3: Examples of (a) full-body images and (b) head images of participants facing each direction, and visualization of their HOG features.

test set. Together, we call this challenging dataset the Human Orientation DataSet, or HumODS.

C. Multiple kernel learning

As explained in Sec. III-B, we have eight total directions: north, east, south, west, northeast, southeast, southwest, and northwest. Dalal and Triggs used the histogram of oriented gradients (HOG) on a large dataset of photos of pedestrians in [14], and HOG features have been successfully used in other identification tasks, as in [15]. We therefore take the HOG features of every photo in our dataset, using the testing and training images detailed in Sec. III-B. Examples of training images and their HOG features are shown in Fig. 3.

Multiple kernel learning (MKL) [16] is able to combine features at different levels in a well founded way that learns to incorporate a predefined set of SVM kernels automatically. It aims at removing assumptions of kernel functions and eliminating the burdensome manual parameter tuning in the kernel functions of SVMs. Formally, it defines a convex combination of m kernels. The output function is formulated as follows:

$$s(\boldsymbol{x}) = \sum_{k=1}^{m} \left[\beta_k \left\langle \boldsymbol{w}_k, \Phi_k(\boldsymbol{x})\right\rangle + b_k\right] \qquad (1)$$

where $\Phi_k(\boldsymbol{x})$ maps the feature data $\boldsymbol{x}$ using one of m predefined kernels, with an L1 sparsity constraint. The goal is to learn the mixing coefficients $\boldsymbol{\beta} = (\beta_k)$, along with $\boldsymbol{w} = (\boldsymbol{w}_k)$, $\boldsymbol{b} = (b_k)$, $k = 1, \ldots, m$. The resulting optimization problem becomes:

$$\min_{\boldsymbol{\beta},\boldsymbol{w},\boldsymbol{b},\xi} \frac{1}{2}\Omega(\boldsymbol{\beta}) + C\sum_{i=1}^{N} \xi_i, \text{ s.t. } \forall i : \xi_i = l\left(s(\boldsymbol{x}^{(i)}), y^{(i)}\right) \quad (2)$$

where $(\boldsymbol{x}^{(i)}, y^{(i)}), i = 1, \ldots, N$ are the training data and N is the size of the training set. Specifically, $\boldsymbol{x}^{(i)}$ is a HOG feature vector, with its corresponding training label $y^{(i)} = 1$ for a positive sample and $y^{(i)} = -1$ otherwise.

In Eq. 2, C is the regularization parameter and l is a convex loss function, and $\Omega(\boldsymbol{\beta})$ is an L1 regularization parameter to encourage a sparse $\boldsymbol{\beta}$, so that a small number of kernel functions are selected. This problem can be solved by iteratively optimizing $\boldsymbol{\beta}$ with fixed $\boldsymbol{w}$ and $\boldsymbol{b}$ through linear programming, and optimizing $\boldsymbol{w}$ and $\boldsymbol{b}$ with fixed $\boldsymbol{\beta}$ through a generic SVM solver.

Equations 1 and 2 depict the standard binary classifier. In this work, they are extended to address the multiclass classification problem by one-against-all implementation of binary classifiers.

TABLE I: A quantitative comparison of the performance of the models used.

Images	Kernel(s)	Accuracy	F1-score
body	linear	0.207	0.151
body	Gaussian	0.241	0.216
body	MKL	0.296	0.275
head	linear	0.453	0.453
head	Gaussian	0.485	0.483
head	MKL	**0.515**	**0.508**

TABLE II: Confusion matrix of the best performing model (MKL on head images).

	E	NE	N	NW	W	SW	S	SE
E	24	13	0	0	7	4	0	3
NE	3	39	0	0	0	7	2	0
N	0	2	28	8	0	2	5	8
NW	4	1	2	28	7	0	2	10
W	5	7	3	4	19	24	1	1
SW	2	14	7	1	1	27	2	0
S	2	2	17	2	0	2	24	6
SE	3	0	0	10	2	0	5	37

IV. Experiments and Results

We used MKL to classify direction orientation over the body images and, separately, the head images. We trained three models on each image set – a linear kernel SVM, a RBF kernel SVM, and a MKL model. For the single kernel approaches, the regularization parameter C was optimized using a 3-fold cross validation. We selected a fixed Gaussian kernel (σ=0.5) for the RBF models. The MKL approach automatically selected kernels from a list of Gaussian RBF kernels (σ = 0.5, 1, 2, 5, 7, 10, 12, 15, 17, 20) and polynomial kernels (degree = 1, 2, 3). Its regularization parameter C was 1. The classification performance was measured with accuracy and F_1 score. Accuracy is defined in terms of true positives, TP (positive examples labelled correctly), false positives, FP (negative examples incorrectly labelled as positives), true negatives, TN (negative examples labelled correctly) and false negatives, FN (positive examples incorrectly labelled as negatives). Accuracy is determined by the equation

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

The F1 score combines precision ($TP/(TP + FP)$ or the fraction of classes labelled as a label that was correct) and recall ($TP/(TP + FN)$ or the fraction of classes in a label that were actually found) into one number, denoted as

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

Table I compares the performance of these models.

In general, the single overhead images are difficult to classify because of the lack of facial features. The MKL approach outperforms single kernel SVMs, by automatically selecting the best kernels. Particularly, head images are better than full body images for the task of orientation classification. The performance gap between these two image sets is over 20% accuracy. As shown in Fig. 3, in most of the images, children are sitting on the ground, with a variety of body poses. Therefore, including the body region may introduce more noise than useful features.

Table II shows the confusion matrix of the classification. It can be seen that misclassification often happens between opposite directions or adjacent directions. Fig. 4 presents success and failure examples of the test images, which include correct classifications (in blue boxes), misclassified adjacent directions (in yellow boxes), and other misclassified examples (in red boxes). Note that the test images are of different subjects from the training images. The model may fail when the test sample has a different hair color or style that is not seen in the training set.

V. Conclusions and Future Work

Ultimately, our goal is to classify a child's orientation in any given frame of an overhead video during a human-robot interaction experiment. While our current approach can be improved, we have demonstrated that overhead orientation is a tractable problem even with a single overhead camera, enabling us to begin categorizing more directions and orientations than previous work with a simpler set-up. We showed that MKL performs better than single kernels, and that using smaller person features (i.e. heads) performs better than larger, possibly noisier person features (i.e. whole bodies).

We have several clear-cut next steps: training our model with more sample data and thus generating a larger dataset, applying our model to a frame-by-frame replay of video footage, taking advantage of the person orientation from previous frames, and comparing our MKL approach with deep learning approaches.

As mentioned earlier, we are currently running human-robot interaction experiments, as we are restrained in our data collection by the schedules of the parents of our participants. In the near future there will be still more data to add to our dataset. If there are some pertinent features that do not show up frequently in our dataset, such as particular hairstyles or body types, introducing more child data into the HumODS dataset may help the MKL performance. Our participant pool includes boys and girls of varying ethnic backgrounds, so there will be natural deviations in the data for MKL to train with. We expect the addition of future experiments to our existing database will improve the performance of all models.

With a more robust dataset and model, the next step will be to apply our model to video footage of experiments, frame-by-frame. Running the model on an entire video will therefore give our orientation estimation another piece of information for every frame but the first frame – we will have a history of orientation thus be able to weight that direction as more likely in the event of a tie or uncertain classification.

Lastly, we will turn to deep learning on this dataset. It may be that details that don't show up in HOG features do appear

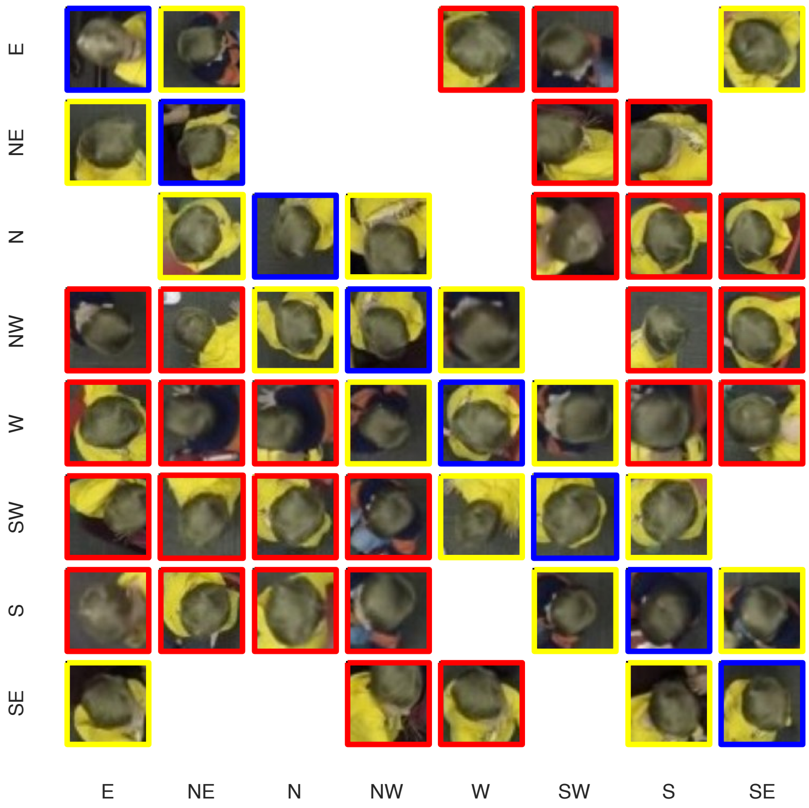


Fig. 4: Qualitative visual representation of correctly and incorrectly labelled images. Blue boxes are correct classifications, yellow boxes are misclassified adjacent directions, and red boxes are the other misclassified examples.

after training with neural networks; there may be abstract notions of hair styles or shoulder positions that give orientation information that are confounding and noisy for our current model. The noisy body orientations especially that degenerate our MKL models may be quite useful in a neural network.

Acknowledgment

The authors thank the NSF for support under grant IIP-1439728, and the anonymous reviewers for their comments.

References

[1] B. Scassellati, “How social robots will help us to diagnose, treat, and understand autism,” *Robot. Res.*, pp. 552–563, 2007.

[2] K. Asada, Y. Tojo, H. Osanai, A. Saito, T. Hasegawa, and S. Kumagaya, “Reduced personal space in individuals with autism spectrum disorder,” *PLoS One*, vol. 11, no. 1, 2016.

[3] L. Zwaigenbaum, S. Bryson, T. Rogers, W. Roberts, J. Brian, and P. Szatmari, “Behavioral manifestations of autism in the first year of life,” *Int. J. Dev. Neurosci.*, vol. 23, no. 2-3 SPEC. ISS., pp. 143–152, 2005.

[4] L. K. Koegel, R. L. Koegel, K. Ashbaugh, and J. Bradshaw, “The importance of early identification and intervention for children with or at risk for autism spectrum disorders.” *Int. J. Speech. Lang. Pathol.*, vol. 16, no. 1, pp. 50–6, 2014.

[5] J. Hashemi, T. V. Spina, M. Tepper, A. Esler, V. Morellas, N. Papanikolopoulos, and G. Sapiro, “Computer vision tools for the non-invasive assessment of autism-related behavioral markers,” *2012 IEEE Int. Conf. Dev. Learn. Epigenetic Robot. ICDL 2012*, pp. 1–33, 2012.

[6] J. Bidwell, I. a. Essa, A. Rozga, and G. D. Abowd, “Measuring child visual attention using markerless head tracking from color and depth sensing cameras,” *Proc. 16th Int. Conf. Multimodal Interact. - ICMI ’14*, pp. 447–454, 2014.

[7] J. Hashemi, M. Tepper, T. Vallin Spina, A. Esler, V. Morellas, N. Papanikolopoulos, H. Egger, G. Dawson, and G. Sapiro, “Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants.” *Autism Res. Treat.*, 2014.

[8] J. Fasching, N. Walczak, V. Morellas, and N. Papanikolopoulos, “Classification of motor stereotypies in video,” *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 4894–4900, 2015.

[9] R. Mead, A. Atrash, and M. J. Matarić, “Automated proxemic feature extraction and behavior recognition: applications in human-robot interaction,” *Int. J. Soc. Robot.*, vol. 5, no. 3, pp. 367–378, 2013.

[10] “TOBII Pro.” [Online]. Available: www.tobiipro.com/

[11] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *Fg*, vol. 23, no. 6, pp. 484–498, 1998.

[12] D. Cristinacce and T. Cootes, “Feature detection and tracking with constrained local models,” *BMVC 2006 - Proc. Br. Mach. Vis. Conf. 2006*, pp. 929–938, 2006.

[13] R. Sivalingam, A. Cherian, J. Fasching, N. Walczak, N. Bird, V. Morellas, B. Murphy, K. Cullen, K. Lim, G. Sapiro, and N. Papanikolopoulos, “A multi-sensor visual tracking system for behavior monitoring of at-risk children,” in *Robot. Autom. (ICRA), 2012 IEEE Int. Conf.* IEEE, 2012, pp. 1345–1350.

[14] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, pp. 886–893, 2005.

[15] O. Lanihun, B. Tiddeman, E. Tuci, and P. Shaw, “Improving active vision system categorization capability through histogram of oriented gradients,” *Towar. Auton. Robot. Syst.*, pp. 143–148, 2015.

[16] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, “Choosing multiple parameters for support vector machines,” *Machine Learning*, vol. 46, no. 1-3, pp. 131–159, 2002.