

Camera-based estimation of student's attention in class

THÈSE N° 6745 (2015)

PRÉSENTÉE LE 23 OCTOBRE 2015

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

LABORATOIRE D'ERGONOMIE ÉDUCATIVE

PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Mirko RACA

acceptée sur proposition du jury:

Dr R. Boulic, président du jury
Prof. P. Dillenbourg, directeur de thèse
Prof. X. Ochoa, rapporteur
Prof. D. Gasevic, rapporteur
Prof. D. Gatica-Perez, rapporteur



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

The good life is one inspired by love and guided by knowledge.
Bertrand Russell

Posvećeno mojim roditeljima – Ljubomiru i Biljani.
(Dedicated to my parents – Ljubomir and Biljana.)

Acknowledgements

Five years of life can not be easily summed up in a page-and-a-half of text. Sometimes the best office discussions were not professional at all, and that was the exact reason why they were important for me, and indirectly for this thesis. With that, I can say that my stay at EPFL was a fantastic personal and professional experience, and I will try to list people who made the biggest impact on me during this period. The complete list would be too long, and for all of those omitted, I beg forgiveness.

It is safe to say that this thesis would never happen if it was not for professor **Pierre Dillenbourg**. His aid reached beyond academic guidance, and I can attest that a number of people (me included) will leave CHILI laboratory not only as better scientists, but also as better persons because of him. He is the pace-maker of the longest run.

Half of the experiments in this publication were made possible by the support of **Roland Tormey**, and I have long lost the count of good discussions and insights that were made in our conversations. There were a many occasions on which I wished I could take notes during our lunch chats, and the fragment that I did remember certainly made it into this work.

I remain thankful to the entirety of the CRAFT/CHILI personnel for their discussions, friendship and support. Over time, I changed three post-doctoral advisors - **Jan Blom**, **Guillaume Zufferey** and **Łukasz Kidziński**. Each of them contributed to my work with their unique perspective on the problem, and I am grateful for our discussions. **Patrick Jermann** remains the expert of practical statistics, a title that is connected with the fact that he found errors in my analysis more often than I was comfortable with. I am thankful to all the people who made my stay in office and outside of it more joyful - **Kshitij**, **Hamed Alavi**, **Quentin Bonnard**, **Mina Shirvani Boroujeni**, **Sebastien Cuendet**, **Jessica Dehler Zufferey**, **Julia Fink**, **Alexis David Jacq**, **Severin Lemaignan**, **Nan Li**, **Lorenzo Lucignano**, **Nikos Maris**, **Andrea Mazzei**, **Ayberk Ozgur**, **Luis Pablo Prieto Santos**, **Flaviu Roman**, **Beat Schwendimann** and **Himanshu Verma**. All of us were backed up by **Florence Colomb**, who kept a steady rhythm during the whole “gig” (a band is only as good as its drummer).

From my stay in the Computer Vision laboratory, I would like to especially thank **Aurelien Lucchi**, **Horesh Ben Shitrit** and **Vincent Lepetit**. Their explanations and help were invaluable, and they have set a high bar of what it means to be a good computer scientist which I am still

Acknowledgements

chasing.

My stay at EPFL also gave me the opportunity to become a proud member of the band *Uncontrollables*, which apart from terrible music also produced wonderful friendships. I would like to thank **Zlatko Emeđi**, **Sean Costello**, **Simone Formentin**, **Alan Bock** and **Maja Đukic Pjanić** for having patience for my mistakes, and keeping the music alive. In extension, the band's expanded composition has to include **Nataša** with **Lenka**, and **Giorgia** with **Alice**, who made all the difference. Also, a special mention is in order to the original lunch-group of "*Žderilnici*" – **Andreas**, **Aleks**, **Bapra**, **Dražen**, **Gvero** and **Peda**.

Among all friends I gained during my stay in Lausanne, I will miss the most **Orest Kuzyk** who stays forever in my memories.

To my extended family Ilievski – **Aco**, **Jasmina** and **Nataša** – I thank for their encouragements in the last stages, they were greatly appreciated.

Three people who did everything they could and more to help me on every step of the way – my father **Ljubomir** who believed in me more than I ever did, my mother **Biljana** who was my wit when I had none, and my brother **Igor** who is making me re-evaluate everything I do every year anew. I dedicate this thesis to them.

I can only mention **Dimitrije** and **Sonja**, because the love and support I received from each of them (in their own way) are beyond this text. You give meaning to my efforts and make me await each new day with anticipation. Our journey is just beginning.

Lausanne, 6 October 2015

M. R.

Abstract

TWO essential elements of classroom lecturing are the teacher and the students. This human core can easily be lost in the overwhelming list of technological supplements aimed at improving the teaching/learning experience. We start from the question of whether we can formulate a technological intervention around the human connection, and find indicators which would tell us when the teacher is not reaching the audience.

Our approach is based on principles of unobtrusive measurements and social signal processing. Our assumption is that students with different levels of attention will display different non-verbal behaviour during the lecture. Inspired by information theory, we formulated a theoretical background for our assumptions around the idea of synchronization between the sender and receiver, and between several receivers focused on the same sender. Based on this foundation we present a novel set of behaviour metrics as the main contribution.

By using a camera-based system to observe lectures, we recorded an extensive dataset in order to verify our assumptions. In our first study on motion, we found that differences in attention are manifested on the level of audience movement synchronization. We formulated the measure of “motion lag” based on the idea that attentive students would have a common behaviour pattern.

For our second set of metrics we explored ways to substitute intrusive eye-tracking equipment in order to record gaze information of the entire audience. To achieve this we conducted an experiment on the relationship between head orientation and gaze direction. Based on acquired results we formulated an improved model of gaze uncertainty than the ones currently used in similar studies.

In combination with improvements on head detection and pose estimation, we extracted measures of audience head and gaze behaviour from our remote recording system. From the collected data we found that synchronization between student’s head orientation and teacher’s motion serves as a reliable indicator of the attentiveness of students. To illustrate the predictive power of our features, a supervised-learning model was trained achieving satisfactory results at predicting student’s attention.

Key words: computer vision, non-verbal behaviour, social signals, motion synchronization, gaze usage, head motion, student’s attention, classroom entropy

Résumé

Les deux éléments essentiels de l'enseignement en classe sont l'enseignant et les étudiants. Ce point peut facilement être perdu de vue dans la liste imposante des moyens technologiques visant à améliorer l'expérience d'enseignement et/ou d'apprentissage. Notre point de départ sera de voir si l'on peut faire intervenir la technologie autour de la connexion humaine et de trouver des indicateurs nous signalant quand l'enseignant n'atteint pas l'attention de son public.

L'approche est basée sur le principe de prise de mesures non-intrusives et de traitement de signaux sociaux. Notre hypothèse est que les étudiants ayant divers niveaux d'attention vont exprimer différents comportements non-verbaux. Inspirés par la théorie de l'information, nous avons développé un environnement théorique pour notre hypothèse autour de l'idée de synchronisation entre l'émetteur et le récepteur, ainsi qu'entre plusieurs récepteurs concentrés sur le même émetteur. En nous appuyant sur ce fondement, nous présentons une série de mesures du comportement novatrice comme contribution principale.

En utilisant un système de caméras pour observer les cours, nous avons enregistré un large éventail de situations afin de vérifier notre hypothèse. Dans notre première étude sur le mouvement, nous avons découvert que la différence d'attention se manifeste à un niveau aussi simple que la synchronisation des mouvements du public. Nous avons énoncé la notion de «décalage de mouvement» (motion lag) se basant sur l'idée que les étudiants étant attentifs auraient un comportement commun.

Pour notre seconde série de mesures, nous étudions la manière de remplacer l'équipement d'oculométrie intrusif afin d'enregistrer les informations sur le regard de tout un public. Pour y parvenir nous avons mené une expérience sur la relation entre l'orientation de la tête et la direction du regard. Se basant sur les résultats obtenus nous avons créé un modèle de calcul de l'angle du regard amélioré par rapport à ceux utilisés actuellement dans des études similaires.

En combinaison avec les améliorations de la détection des mouvements de la tête et des évaluations de la posture, nous avons extrait des mesures du comportement de la tête et du regard du public à l'aide notre système d'enregistrement installé en retrait. D'après les données recueillies, nous avons découvert que la synchronisation entre l'orientation de la tête des étudiants et des mouvements de l'enseignant donne des indices fiables quant à l'attention

Résumé

des étudiants. Pour démontrer les possibilités de prédictions de notre système, les données ont été soumises à un modèle d'apprentissage automatique qui a obtenu des résultats satisfaisant pour prédire l'attention des étudiants.

Mots clé : vision numérique, comportement non-verbal, signaux sociaux, synchronisation de mouvements, utilisation du regard, mouvement de la tête, attention de l'étudiant, entropie de la classe.

Contents

Acknowledgements	i
Abstract	iii
List of figures	xi
List of tables	xix
1 Introduction	1
1.1 Motivation	1
1.2 Research objectives	3
1.3 Contributions	3
1.4 Organization of this thesis	4
2 Related work	5
2.1 On attention	5
2.2 Signal-to-noise in human communication	7
2.2.1 Digitalization of human interactions	9
2.3 Non-verbal communication	12
2.3.1 Posture	14
2.3.2 Hand and body gestures	14
2.3.3 Head gestures	15
2.3.4 Facial expressions	16
2.4 Visual focus of attention (VFoA)	17
2.5 Head pose estimation	18
2.5.1 Problem formulation	18
2.5.2 Technical approaches	20
2.5.3 VFoA in small group situations	23
2.6 Usage of measurements	25
2.6.1 Individuals	25
2.6.2 Groups and interactions	27
2.7 Classrooms	28
2.7.1 The shape of the classroom	29
2.7.2 Attention in learning	30

Contents

2.8	Conclusion	31
3	Channels and signals of classroom interactions: an informational view	33
3.1	Elements of classroom communication	34
3.1.1	Sources of messages	35
3.1.2	Signals	36
3.1.3	Channels	36
3.1.4	Synchronization	37
3.2	Spatial dimension	38
3.3	Interpreting the classroom	40
4	Recorded attention and classroom behaviour samples	41
4.1	Capturing devices	41
4.1.1	Post processing	44
4.2	Sample statistics	45
4.3	Capturing changes in attention	46
4.3.1	In-class questionnaires	47
4.3.2	Post-class questionnaires	48
4.4	Data in the questionnaires	49
4.4.1	Differences between in-class and post-class format	49
4.4.2	Results captured in the questionnaire	51
4.5	Mobility of students	56
4.6	Interviews with the students	57
4.7	Conclusion	60
5	Movement in class	63
5.1	Method	64
5.1.1	Optical flow	65
5.1.2	Distribution of tracking features	68
5.1.3	Outlier elimination	69
5.1.4	Motion tracks and assignment	69
5.2	Motion Analysis	71
5.2.1	Motion normalization	72
5.2.2	Observable and synchronized movement	73
5.2.3	Motion synchronization	73
5.3	Research questions	75
5.4	Results	76
5.5	Conclusion	78
6	Participation of head and eyes	81
6.1	Previous work	81
6.2	Methodology	83
6.2.1	Real-world usage	84

6.2.2	Controlled experiment	84
6.3	Research questions	89
6.4	Results	90
6.4.1	Real-world usage	90
6.4.2	Controlled experiment	91
6.5	Gaze limits from head pose	94
6.6	Conclusion	95
7	Gaze in the classroom	101
7.1	Data extraction	102
7.1.1	Manually annotated data	102
7.1.2	Head detection with Part-Based Model	102
7.1.3	Pose estimation	107
7.1.4	Head pose tracking	110
7.1.5	Teacher tracking	115
7.2	Defining behaviour measures	117
7.2.1	Individual measures	117
7.2.2	Modelling behaviour over time	118
7.2.3	Influence of distance on field of view	121
7.3	Research questions	122
7.4	Results	122
7.4.1	Individual measurement analysis	122
7.4.2	Teacher's movement analysis	124
7.4.3	Connection between student's gaze and teacher's position	125
7.5	Conclusion	129
8	Conclusions	131
8.1	Summary	131
8.1.1	Results	132
8.1.2	Contributions	133
8.2	Limitations	134
8.3	Future work	135
	Bibliography	153
	Curriculum Vitae	155

List of Figures

2.1	a) Camera view and b) schematic diagram of the IBM smart room, taken from (Mostefa et al., 2007). The diagram shows the locations of 9 cameras and 152 microphones used for data collection, used for capturing 4-person meetings. .	10
2.2	Behavioural cues and social signals, taken from Vinciarelli et al. (2008). The scene, even with the minimal amount of information is clearly interpretable through the cues of gestures and body posture.	11
2.3	Equipment for capturing human interactions. Shown in the image are variety of interview cameras, web-cameras, depth cameras, portable eye-tracker, portable EEG, mobile devices (microphone, accelerometer) and input devices.	13
2.4	Detecting emotions in drivers. Taken from Gao et al. (2014).	17
2.5	The axes of head rotation are horizontal rotation (yaw or pan, θ) , vertical rotation (pitch, ϕ) and sideways rotation (tilt or roll, χ) . Image taken from (Murphy-Chutorian and Trivedi, 2009).	19
2.6	CHIL dataset samples (Waibel et al., 2009). Different scenarios were included from a) single person in the room, b) to meeting scenarios with c) dense coverage of the meeting space from several camera viewpoints. Images taken from (Voit et al., 2008).	23
2.7	Augmented Multi-party Interaction (AMI) dataset (Carletta et al., 2006). a) Overview of the space, b) one of the dedicated cameras for two of four participants and c) general arrangement of the space. Images taken from (Odobez and Ba, 2007; Ba et al., 2009).	24
3.1	Signal classification based on modality and structural base.	36
3.2	Organization of classroom zones and units of measurements, top of the image represents the front of the classroom. v,h - vertical and horizontal spacing between students is 1 <i>unit of distance (uod)</i> . s - between-row spacing, 1 <i>uod</i> . d - distance between the professor (center-front of the classroom) and the analysed student. Light-blue zone represents the visible students for student at the location 3rd row, 4th seat. Darker-blue rectangle around the student represents his immediate neighbourhood.	39
4.1	Equipment deployed during the experiment in classroom B.	42

List of Figures

4.2	Layouts of classrooms a) A and a) B. Camera positions and orientations are designated with bright-blue boxes on the sides of the lecture's area (as used in the head-detection recordings). Gray tables represent the seats which were not used during the lecture.	43
4.3	The a) in-class and b) post-class questionnaire used during our experiments. .	47
4.4	Measurements compared across different questionnaire types. In case of <i>a)</i> attention and <i>b)</i> classroom attention no statistically significant differences were observed. In cases of <i>c)</i> teacher energy and <i>d)</i> material importance, ANOVA showed significant differences between questionnaires (in both cases $p < 0.001$). .	50
4.5	Comparing a) material importance and b) attention captured over the four periods with the in-class and post-class questionnaire format.	51
4.6	Aggregated attention samples of all students in the study ($\mu = 6.79$, $\sigma = 2.07$). Colours designate the attention level labels associated with each reported level (Number of samples: low (159 samples, 14.79%), medium (464 samples, 43.16%) and high (454 samples, 42.23%))	52
4.7	Attentions of all classes, with different student populations visualized in different colours. Bars are presented in the order in which the data was captured. Populations 1, 4 and populations 2, 3 are associated with the same lecturer. . .	52
4.8	Changes of attention over <i>a)</i> the 4 sampled times (1075 samples); <i>b)</i> the distance from the teacher (225 samples, $b = -0.19$, $r^2 = 0.02$, $p = 0.01$).	53
4.9	Attention transition values for <i>a)</i> individual levels displayed, increase of attention is represented in the upper triangle of the matrix, while the blue line represent maintaining the same level of attention; <i>b)</i> transitions between discrete levels of attention (<i>low/medium/high</i>) declared in the Section 4.4.2.1.	54
4.10	Relationship between <i>a)</i> prior interest and mean reported attention (sample size = 97, $b = 0.48$, $r^2 = 0.22$, $p < 0.01$), <i>b)</i> prior knowledge and mean reported attention (sample size = 97, $b = 0.26$, $r^2 = 0.06$, $p < 0.01$), and <i>c)</i> mean reported attention and results in the post score (sample size = 86, $b = 0.02$, $r^2 = 0.03$, $p = 0.056$).	54
4.11	Relationship between attention and <i>a)</i> classroom attention $r(1026) = 0.43$ ($p < .0001$), <i>b)</i> teacher's energy $r(1056) = 0.25$ ($p < .0001$), and <i>c)</i> material importance $r(1060) = 0.41$ ($p < .0001$). To display the density of many overlapping points, Gaussian noise ($\sigma = 0.2$) was added to the location of the points after the linear model was fitted and individual points were made semi-transparent.	55
4.12	Percentage of reports of each activity per attention level in <i>a)</i> Population 1 and <i>b)</i> Population 2. Number of reported instances was normalized by the total number of instances on that attention level to produce the percentages.	56
4.13	Histogram of student location changes.	57
4.14	Modality of focus emphasized by students <i>a)</i> based on the number of attributes used divided by their modality, <i>b)</i> visual target of attention mentioned.	59

4.15 Perception of lecturer based on the attributes used to describe his teaching <i>a)</i> divided by the type of attribute (percentage represents the ratio of all attributes used), <i>b)</i> list of attributes and their usage frequency (colours denote the type of attribute).	60
4.16 Perception of the experiment by the interviewed students. <i>a)</i> Distribution of opinion about the overall experiment participation. <i>b)</i> Intrusiveness of the experiments in terms of how many subjects agreed with a certain statement. . .	61
5.1 Annotated student regions and overlaps. a) Rectangles represent the annotated regions connected with each student ID (different colours of rectangle's borders used for better visualization of occurring overlaps). b) Each edge of the graph represents an overlap between two regions and potential ambiguity for motion assignment. Direction of the edges is oriented towards the overlapped (occluded) student.	65
5.2 Algorithm used for motion extraction and assignment. Overview of motion detection is given in Section 5.1.1. Distribution of tracking points is described in Section 5.1.2. Outlier elimination is detailed in Section 5.1.3. Motion tracks are described in Section 5.1.4.	66
5.3 Illustration of the optical flow. Individual image frames 1-3 are shown in the top row, while the difference between two respective frames is shown in the bottom. Optical flow is represent by the arrows representing motion vectors, shown for each pixel in the bottom row. Image copyright by Florian Raudies, Licence: CC Attribution 3.0.	67
5.4 Visualization of the pyramid down-sampling, 0th level representing the original image size, and each next level scaling down image size by half. Image sizes under the layer label are given for illustration purposes.	67
5.5 Distribution of tracking points for the optical flow. Each red dot represent a point for measuring a motion vector. Notice the difference of densities for the students in front and in the back. In regions with no good tracking features detected, tracking points are arranged in a equally-spaced grid (visible in the lower right corner).	68
5.6 Motion detection and grouping. a) Each annotated region is associated with a student ID. b) Marked student areas and illustrated centres of 2D Gaussian distributions which model the probability of motion belonging to a student. c) Individual motion vectors shown as purple arrows. d) Motion vectors grouped into motion tracks (represented as jagged lines) which can be assigned to an individual.	70

List of Figures

5.7	Visualization of motion intensity for a single person (green) over the class mean (grey). Horizontal axes represents the time, vertical the relative intensity of motion for the given person. Vertical markers represent the annotated classroom events, most notably - the 4 paired red markings represent the moments of start and end of the questionnaire fill-out. Horizontal red line shows the minimal intensity of motion considered an “observable movement”.	72
5.8	Example of co-movement on a movement intensity graphic of two persons. The picture is a snippet of a motion-visualization as shown on Figure 5.7. We are displaying motion intensity of two persons overlayed over each other. Person 2 shifted hers seating position (blue line), 2 seconds later, neighbouring Person 1 (marked in green) also started re-adjusting herself.	73
5.9	Synchronized movement. a) Co-movement matrix of Person A and B over a period of 12 seconds (6 time steps). Perfect synchronization is represented by the diagonal of the matrix, marked with red squares. $< \pm 4$ second synchronization is represented with blue cells and weak synchronization ($< \pm 6$ seconds) is marked with green cells. Periods which were too far apart to be considered are grayed-out. b) Co-movement timeline, considered from the perspective of Person B. The figure shows the same values as the co-movement matrix, aligned on the diagonal cells of the matrix (red squares). Transparent sections are not present in the example matrix.	75
5.10	Difference in the number of observed synchronized moments between pairs of neighbouring students and other pairs of students. Values shown for a) Class #1 and b) Class #2. Exact values shown in Table 5.1. Boxplots show the median value, the edges of the boxes are the 25th and 75th percentile and whiskers depict the 90% of the sample.	76
5.11	Correlation between distance from teacher and motion intensity in Class 2; Kendall correlation $\tau = -0.284$ ($p = 0.03$)	77
5.12	Motion lag compared with the mean level of attention. a) Class 2 shows a Kendall correlation $\tau = -0.259$ ($p = 0.06$). b) Both classes combined show weaker correlation, Pearson's $r = -0.011$, ($p = 0.09$). Colours of data points represent Class 1 and 2 samples. Gray regions of fitted lines in both cases represent the 95% confidence interval.	79
6.1	Portable eye-trackers. Two of the popular brands are the a) Tobii Glasses and b) SMI portable eye-tracker, shown here with the SMI Optical Head Tracking module. Images copyrighted by the mentioned companies.	83
6.2	Undistorted view angles of the eye-tracker in degrees, shown for every location of the recorded view-field. Centre is marked with the white horizontal and vertical lines to display the horizontal and vertical offset.	83

6.3	a) Teacher's view from one of the recorded classes. b) Top-view of the controlled experiment showing the seated student wearing the head-pose tracking markers (chili-tags) on his head. Recording and synchronization devices visible on the right half of the image.	86
6.4	Visualization of the data captured in the controlled experiment. The three streams of data shown are the ground truth of the point angular position (α_{GT} , blue line), horizontal angle of the eyes (α_E , green) and head (α_H , red). The vertical red blocks were the periods when the stimulus was not in the "observed" state. Shorter unobserved periods within a single angular fixation are not visible due to their short duration.	86
6.5	Visual stimuli shown to the subjects during the controlled experiment. a) The stimulus would initially be shown as a yellow cross for easy localization by the subject, b) after the user has clicked the mouse button (indicating that he sees the item), the item would change its shape and colour to green circle to indicate that c) when the item was moving, it would be displayed as the white square to enable easy tracking by the subject.	87
6.6	State transitions used in the controlled experiment. Dashed lines represent automatic transitions, and full line represents the effect of the user's input. . . .	88
6.7	Unscripted gaze behaviour captured from the teacher's in class. The line colours represent. subjects 1 and 2, and the black dots show the mean values for the 25%, 50%, 75%, 90% and 95% of observed gaze locations. Exact values are shown in Table 6.3.	91
6.8	a) Heatmap of a single recorded teacher. Note that the centre of the distribution is almost identical to the centre of vision (marked with a white cross-hair). White circles designate (from the centre outwards) the 25%, 50%, 75%, 90% and 95% of all captured gaze locations. Binning factor for the heat-plot were zones of 20x20 pixels. b) Deviating sample of teacher's vision.	92
6.9	Absolute errors in angles for each of the fixation locations. Outliers of greater magnitude were present, but were cut from the graph for clarity of the main body of data.	92
6.10	Visualized level of head participation. Red line represents the amplitude (absolute value) of head rotation over the sampled range of horizontal angles modelled by Equation 6.3. As a reference, green line visualizes the hypothetical 100% of head participation. Box-plots show the variance of collected head rotations samples (boxes show the 25-75 percentiles, and black lines show the mean values). . . .	93
6.11	Changes in head pose relative to the initial fixation. a) Shown for the two transition conditions aggregated. b) Changes in head pose shown for each angle, transition conditions represented with colours.	94

List of Figures

6.12	Field of view limits modelled from the head orientation. a) Shows all the components of the final function overlaid. Linear relationship from the Equation 6.4 is modified with imposing the minimal uncertainty of $\pm 18.61^\circ$ and limiting the extreme values to $\pm 130^\circ$ modelling the extreme head and eye rotation. b) The final function visualized.	97
6.13	Progression of head-pose in absolute values (positive is away from the centre, negative values are towards the centre) for different fixation angles. Travel condition and angles are indicated in the title of the graph. The angle was set to zero on the moment the stimulus was first indicated as observed, and the relative progression of head pose was tracked over time from there. Prominent blue line indicates fitted linear model on all data points observed.	98
6.14	Progression of head-pose in absolute values (positive is away from the centre, negative values are towards the centre) for different fixation angles. Travel condition and angles are indicated in the title of the graph. The angle was set to zero on the moment the stimulus was first indicated as observed, and the relative progression of head pose was tracked over time from there. Prominent blue line indicates fitted linear model on all data points observed.	99
6.15	Progression of head-pose in absolute values (positive is away from the centre, negative values are towards the centre) for different fixation angles. Travel condition and angles are indicated in the title of the graph. The angle was set to zero on the moment the stimulus was first indicated as observed, and the relative progression of head pose was tracked over time from there. Prominent blue line indicates fitted linear model on all data points observed.	100
7.1	Part-based models a) Visualization of the tree-of-parts for the detection of human figure; b) HoG signatures for each of the parts visualized; c) model for face detection; d) example of DPM for car detection; e) example of pose detection/estimation; f) example of face detection with visualized sub-parts. Images taken from (Felzenszwalb et al., 2010; Yang and Ramanan, 2011; Zhu and Ramanan, 2012).	103
7.2	Mixtures of the original part-based model for detecting faces. A total of 13 mixtures was used by Zhu and Ramanan (2012) for modelling different yaw angles, acquired from the MultiPIE dataset (Gross et al., 2010).	105
7.3	a) Facial landmarks of the AFLW dataset. We used only a subset of all features. b) Example of the training image with correctly placed detections c-f) First four mixtures of the final 7-mixture model. Red lines represent the deformation-tree structures of the facial model, with the HoG visualizations of each part in the background. Angles represented are 0° , 30° , 50° and 70° of yaw. Images a , b are copyright of (Koestinger et al., 2011).	105
7.4	Illustration of the format of images used from Prima Pointing'04 dataset for testing.	107
7.5	Percent of correctly placed detections by the PBD tested on the Pointing'04 dataset. Overall detection score is 56.49%.	108

7.6	Average angular error per head pose (horizontal and vertical angles) a) Original parts-based detector (mixture classification) and b) rbf regression based on 4 features. We accumulated the vertical error to produce the total error per yaw angle for c) original parts-based detector and d) rbf regression.	109
7.7	Problematic situation for detection assignment. a) Both faces are visible and correctly assigned to persons. b) One face is lost, with the other detection being double-assigned (shown here) or miss-assigned.	110
7.8	Dense groups of detections around typical head locations (each dot represents the centre of the detection). Colours represent different GMM mixtures fitted to each cluster of detections. Beside associating the clusters with student ID's, orange cluster in the lower centre would be marked as "outlier" and bright green cluster scattered over several students would be marked as "invalid".	111
7.9	Details of the head tracking algorithm.	114
7.10	Examples of head detections during a) lecture and b) recess.	115
7.11	Example of the video recording of the teacher and visualized tracking region of the TLD tracker (Kalal et al., 2012). The entire width of the front wall was captured, but the image was truncated for display.	116
7.12	Teacher's location coordinates. We display the 0.0 - 1.0 scale in front of the projection area, but note that the coordinates are not truncated to that range. We also display the bins used for positional histograms later shown. Bins are marked as blue and red rectangles in order to differentiate between the "projection" and "outer" zone. Bin width is equivalent to approximately half of body width. . . .	116
7.13	Visualizations of the gaze modelling methods used. Red dot illustrates the student, with blue arrow line showing the centre of view-field and green triangle depicting the field of view. Blue dot represents the teacher's position in the front of the classroom. We illustrate measures used in each method and the output used below the sketch of the classroom.	120
7.14	Shape of Gaussian probabilities modelling gaze probability for the student located in the centre of "projection zone" ($x=0.5$) sitting in different rows. As a reference, vertical lines represent the edges of "projection zone" (0.0 - 1.0 span in the normalized positional coordinates). Models are given for a) Classroom A and b) Classroom B.	121
7.15	Change in normalized head travel correlated to the change in attention. Red line represents the linear fit. Pearson's $r(204) = 0.21$ $p = 0.0011$. Gaussian noise added for the purpose of visualizing the samples without overlap after the linear fit.	123
7.16	Visualization of teacher's tracking. Horizontal axis represent time (vertical lines designate 10 minute intervals). Coloured regions are used to represent the "projection" (blue) and "outer" (red) zone of the teacher's area (as explained in Figure 7.12).	124

List of Figures

7.17 Motion and standing positions of recorded teachers. **a)** Percentage of time spent in each of the positional bins in all recordings. **b)** Percentage of all detections classified as “moving” per bin. 125

7.18 Relationship between different models of gaze and reported attention of students. **a)** Step function, **b)** T-H correlation, **c)** T-H distance between gaze projection and teacher’s position, **d)** Mean (teacher’s) Position Predictability (MPP) - how much does the gaze predict the location of the teacher and **e)** normalized version of the same measurement (nMPP) - values for each student normalized to the range of 0.0-1.0. 126

List of Tables

2.1	Mapping of social cues to social behaviour and technologies used for analysing them. Taken from Vinciarelli et al. (2009)	13
4.1	Basic information about analysed classes. The table shows the number of class/population; number of recorded sessions; number of cameras used (for students, teacher and eye-tracker); size of the population; mean attendance (and variance); number of in-class/post-class questionnaires used; percentage of female population; shape of the classroom (rows/seats in row); number of interviews conducted; the goal of the data collection. * - Number of seats in this classroom varied in every row, mean value displayed.	45
4.2	Parameters collected with the questionnaires, with the number of samples and classes captured. Brackets beside the parameter group name indicate the questionnaire format which was used for that group.	48
4.3	Number of times student showed-up in class during the experiment.	56
4.4	Results of mobility of students.	57
5.1	Average number of synchronized moments between immediate neighbours and other pairs, and results of the t-tests.	77
6.1	Used combinations of horizontal angle, type of travel (S - smooth pursuit, J - jump) and duration of stay in seconds.	85
6.2	Division of angles used between the stay duration and transition type.	85
6.3	Upper angular limits for percentiles of accumulated samples of teacher's gaze. We show angular values for each recording and mean angular limits. Visualization of data is shown in Figure 6.8.	90
6.4	Number of samples collected for the positive and negative values of each horizontal angle.	91
7.1	Details on the training split for different mixtures (modelling different yaw angles) of AFLW dataset and PBM detector. Mixtures are visualized in Figure 7.3.	106
7.2	Comparison of different regression features combinations and kernels for facial pose regression	108

List of Tables

7.3	Features used for the behaviour analysis of the individuals. First three features represent general information connected to spatial location and time of the class. Following six features were extracted from our observations of people's behaviour. Final two are the levels of attention which we will try to predict. . .	118
7.4	Classifier scores for predicting "attention labelled". Score given represent the prediction score on the 20% test sample. Parameters of the kernels are abbreviated as c - penalty for the error term; g - gamma.	124
7.5	Result of Spearman's correlation between the temporal measures modelling contact between the gaze of the student and position of the teacher over periods of 10 minutes.	127
7.6	Result of Spearman's correlation (ρ value / p-value) over different sampling time-steps. Values within each sampling step was aggregated into a single measure by doing mean over the data collected within the time period.	128
7.7	Classifier scores for predicting "attention labelled". Score given represent the prediction score on the held-out 20% test sample. Parameters of the kernels are abbreviated as c - penalty for the error term; g - gamma.	129

1 Introduction

AN anecdotal answer to the question “what does a teacher make” is: “a difference”. Good teachers are figures who remain engraved in our lives, and not only because of the sheer amount of time we spent with them. The *good* teachers made a connection with the classroom and used it to change students’ lives for the better. And yet, after around two decades in the school system and dozens of teachers, only a handful of them will stay remembered, usually the ones who leaned toward the extremes – the very good, or the very bad.

1.1 Motivation

As much as the pedagogical training tries, it will never give to the teachers a full “play-book” for handling a classroom. Similarly, scientific publications will go into detailed analysis of the problem and come up with general guidelines (Hattie, 2008; Davis, 2009). In the end, the major source of “quality” (Pirsig, 1999) in teaching will come from the teachers’ understanding of inter-personal relationships with their students. Our work focuses on that element, by exploring indicators for alerting the lecturers when they have lost the attention of the audience.

Even though different learning theories emphasize different aspects of teaching, or different assumptions about learning altogether, we will discuss formal lecturing. During the lecture teacher’s focus is stretched between personal performance, material integration and supporting students. It is easy to see how one of the elements might fall out of sight. The limitations that we are facing in those moments are real, and even biological to some extent. Practice makes it possible to handle them gracefully, but in teaching terms that usually means years of experience, and thus years of potentially mishandled situations.

As much as the traditional answer to the problem was “*time, practice, experience*”, modern fast pace is inclined to try to find short-cuts with technological aids. If any profession deserves to receive all available resources, a high-profile and high-stake activity such as teaching is one. To assume that all teachers need such technical help would be wrong. Michelangelo also managed quite well without Photoshop. Technological aids in this case are there to act as a

safety net or performance enhancer – a lighthouse that signals you away from the rocks, but does not dictate a specific route you should take. In the last several decades we have seen too many examples of tools becoming the centre of attention and suppressing their original purpose (mobile phones and telephoning; watches and time; computers and calculations). While in other areas of human activity this might even count as a positive shift in the paradigm, traditional teaching should stay focused on transmitting knowledge to students.

With detailed profiling that most modern technologies propose, it is important to touch on the privacy issue connected with associating any measure with human behaviour. Our intent is not to develop an automated system for grading students (or teachers, for that matter), but to provide insights which might guide human interaction in order to make, in our scenario, teaching more effective.

For that reason this thesis aims to keep technology in the shadow of human contact – studying it, but not imposing on it. Unlike other areas where new technology was necessary to enable an activity (e.g. distance learning), person-to-person education in the classrooms does not necessarily need to be pushed through a technological funnel. With this, our intervention naturally found its place among the unobtrusive measurements (Webb et al., 1966) and social sensing (Pentland, 2005). We aim to make the classroom environment more reactive and observant of the human interaction in order to enhance it. It also meant that the environment should not interrupt the lecture with confirmation dialogues, and because of this we modelled our system around the idea of trying to see what the teacher is seeing.

Our intervention is still biased in that sense, and focuses on the teaching staff instead on distributing the information evenly to both sides of the classroom. We root this on the organization setup of classical lecturing (also illustrated as the “*sage on the stage*” model), which is still the dominant form of teaching on majority of educational levels (Moore, 1989). By putting the research as close to realistic scenarios as possible, our conclusions tried to escape the domain of scientific curiosity and be applicable in real scenarios. This is also reflected in equipment used for our experiments which consists primarily of web-cameras and consumer-level computing power.

Research on individuals has the benefits of more developed technological solutions, and personalized lessons are not an uncommon idea on how to improve teaching, so the last question is – why focus on groups of students? Aligned with the concepts of Social Signal Processing (Vinciarelli et al., 2009), our opinion is that there is more information in the relationship between people, than in the behaviour of any single individual in the classroom (the whole being grater than the sum of parts). This view allowed us to pay more attention to the ambient information generated by the audience as a whole.

1.2 Research objectives

Based on the grant from Swiss National Science Foundation (SNSF) ProDoc (project PDFMP1 135108), we launch our investigation with the incentive of expanding the use of unobtrusive measurements of student's attention in the classroom scenario. In order to collect information from a large audience with potentially limited verbal participation, we focus on vision-based features connected with human behaviour extracted with the help of Computer Vision algorithms. Given that one of the main cues that we want to explore is student's approximated gaze direction, we also needed to establish a model of the relationship between student's head orientation and gaze direction.

With this, the main research questions of this doctoral thesis are:

- what are the types of measurements that we can acquire by usage of Computer Vision (CV) techniques in a standard university classroom;
- are CV algorithms capable of scaling up to process dozens of persons in a classroom, and what are the pitfalls;
- is there a detectable inter-personal interaction in a non-collaborative scenario such as listening to the lecture;
- to what extent does the position of the head estimate the gaze direction and how to model the visual focus of attention;
- can the attention of students be assessed on the basis of their non-verbal behaviour.

1.3 Contributions

Contributions presented in the following chapters include work which has been published in a number of papers in areas of Learning Analytics (Raca and Dillenbourg, 2013; Raca et al., 2014), Technology Enhanced Learning (Raca et al., 2013), Multimodal Learning Analytics (Raca and Dillenbourg, 2014) and Educational Data Mining (Raca et al., 2015). Apart from the published work, the manuscript also includes currently unpublished results that were produced in the later stages of our experiments, which includes the results of Chapter 6 and later results of Chapter 7.

Main contributions of the thesis are:

- theoretical view of classroom interactions used as a base for exploring non-verbal behaviour as an indicator of attention,
- usability of motion as an indicator of attention through concept of indirect synchronization of audience members,

- improved model of relationship between head and gaze orientation for horizontal angles,
- exploration of audience head behaviour and relationship between the student's gaze orientation and teacher's position.

1.4 Organization of this thesis

We will start by giving the overview of the related work in the following chapter, focusing on the achievements of non-obtrusive approaches for detecting social cues (Social sensing, Pentland and Heibeck (2008)). We will also justify the need for this intervention with a short overview of observed classroom problems from the pedagogical literature.

Chapter 3 will explain the assumed underlying principles behind our metrics. We will describe the *signal propagation / social entropy* theory of human interactions in order to present our view of classroom interactions. This will be used in further chapters to systematize our efforts and indicate directions of future research.

Chapters 4 through 7 will go through our methodological approach, give details about the processing steps. We will highlight the problems we encountered in order to give a comprehensive overview of the data-sanitizing steps, and methods that we used to extract meaningful signals. Finally we will present our findings on the usefulness of measuring motion and head orientation as indicators of students attention, reached by statistical analysis of the data collected with questionnaires and video-recorded measurements.

2 Related work

As Csikszentmihalyi (2014) described it – attention is the psychic energy that brings order to the chaos of our thoughts. It's directing our efforts, and clearing our mental workbench from the non-essentials so that we can strive towards a specific goal. The benefits of attention for learning are clear, but in order to make it an “actionable” information (Da Silva and Agusti-Cullell, 2008), we need to ask a series of sub-questions.

This chapter will start by analysing attention from the neurological perspective. This is by no means our final interest level or the final word on this complex subject, but it will allow us to define useful properties which other, specialize domains have discovered.

We will continue with the identification of uses of attention in human communication, needed for transmitting knowledge from one person to another. We will touch upon the supporting mechanisms which make communication run smoothly (such as grounding (Clark and Brennan, 1991), back-channels (Argyle, 2013)), and how different modalities play a role in them - specifically the non-verbal communication as evidence of attention and understanding.

From the technical side, we will give an overview of different scenarios and technologies which acted on the information from the non-verbal. The focus will be primarily on human gaze and visual focus of attention (VFoA) as our main interest points.

Finally, we will take the discussion to our domain – the classroom. We will emphasize two dimensions – social and physical – to illustrate problems that teachers are facing, and to analyse how technological aids can support teachers without replacing them or imposing on the learning process.

2.1 On attention

It is possible that there has never been a higher demand for a person's attention. Aside from the usual marketing culprits, TV programs designed around commercials and web-ads, the majority of modern interfaces and public media assume that the user is dedicated exclusively

to them. As Heylighen (2004) noticed, as the communication systems become increasingly effective, the bottle-neck they are facing is our attention. The work of Shenk (1998) depicts a bleak image of people living in a “data smog” of large quantities of low-quality information, which in turn causes anxiety, stress, alienation and errors of judgement.

Research on how to manage such overload lead us into considering the value of messages and the “cost of interruption” for determining whether a specific application should occupy our “attentional spotlight” (Horvitz et al., 2003). Among the cues needed for assessing disruption level, the authors considered a number of contextual information sources, such as: sound analysis, gaze tracking, GPS location, time-of-day only to name a few. The amount of needed information gives us another clue about the complexity of human interactions which we take for granted.

The simple term “attention” hides in itself four important neurological precesses, as identified by Knudsen (2007). The most widely known and important one is *working memory*. Long-time recognized as our measurable buffer (e.g. the “ 7 ± 2 items” rule) and bottleneck, working memory is a resource that different factors are fighting about. Two mechanisms which are responsible for the content of the working memory originate from opposing sides – internal and external.

The internal, *top-down sensitivity control* is probably best described by the Horvitz et al. (2003) model of “attentional spotlight”, “endogenous attention” or “deliberative governing of attention” (Roda and Thomas, 2006). It represents the controlled direction of attention to the stimulus which we find interesting, with underlying activations of different neurological pathways needed for the optimal processing. One of the physical illustrations of this is turning our head in the desired direction, to adjust for better reception of signals from the selected source.

The opposite, externally-driven mechanism is called *bottom-up salience filter* (or “exogenous”, “reactive-governed attention”), which is responsible for enhancing the response to stimuli which is infrequent in space and time. This means that anything which is rare (and, historically speaking – potentially dangerous for us) deserves our attention. In modern times, this mechanism has been widely exploited by the notification systems of various computer applications (Roda and Thomas, 2006).

Finally, the principle which governs the whole process is the *competitive selection of stimuli* – among different present stimuli most of our attention will be directed to the single most interesting one, blocking out others (Rapp, 2006). The last principle was discussed in more detail in the work of Posner and Boies (1971), who organized their research of attention around the concepts of *i*) alertness, *ii*) selectivity and *iii*) processing capacity. Experiments carried out by Posner illustrated our inability to focus on multiple tasks at the same time, even if we are directed to do so. Such limit is know in the literature as the “single-channel limitation”. The experiment was directed at executing two simple tasks divided between visual and audio channels – matching displayed letters and responding to a sound stimuli. The conclusion was

that, although one activity did not completely negate the other (people directed to do the two tasks still tried to carry them out properly), the delay in the execution was correlated to how much the two tasks overlapped temporally.

The target of attention does not necessarily lie in the external world. Episodes of *day dreaming* (Lindquist and McLean, 2011), defined as thoughts unrelated to sensory input, direct the attention of the person inwards. During those periods external activities are being neglected and person's own actions are processed by the default network of cortical regions (Mason et al., 2007; Christoff et al., 2009).

The process becomes more complex when we introduce our neurologically explained individual into a social environment. As social animals, human beings learned to adjust their attention in the presence of other people – an effect which we call “social attention” (Birmingham et al., 2008). Practically illustrated, we tend to look where other people are looking. The cues we pick up are primarily taken from the eye position of other persons, but also from the head and body orientations (Langton, 2000) (if the eyes are not clearly visible, the lower-resolution cues become more important). We can show that this mechanism is more than just a cultural convention by the fact that there is a dedicated part of the brain (inferior temporal (IT) cortex) which is responsible for analysing face and gaze information (Ristic and Kingstone, 2005). The same work shows that when an object is perceived as a pair of eyes, it will continue to have a meaningful signal of gaze direction to the person, even if it doesn't serve that purpose.

Once we have directed our attention, we are primed to process the information coming from the selected source, but with the complexity of inter-human communication, this turns out to be a complex task.

2.2 Signal-to-noise in human communication

Social interactions are defined as any verbal or non-verbal behaviour directed toward or elicited by one or many real or imaginary interaction partners (Mast et al., 2015). In everyday life, communication methods are used non-exclusively and often complementary, forming intricate interaction between them. From a research perspective, this raises the complexity of work by several magnitudes, especially for quantitative analysis which prefers clearly defined, isolated events.

Among the myriad of rules shaping conversation, we are guided by linguistic conventions (language grammar), conversational conventions (e.g. turn-taking) and domain-specific rules (such as lecture participation), just to name a few. But in order to transmit the information successfully, other underlying mechanisms emerge. An example of this is the principle of *grounding* which Clark and Brennan (1991) define as “the coordination of the process of communication”. The principle explains that information propagation is presented in two phases *i)* presentation and *ii)* acceptance. In order for all participants to successfully follow

a common task (e.g. a conversation), “common ground” must be periodically established. Those synchronization check-points are what the authors call the grounding process, and their format have evolved to be very subtle, in order not to disrupt the main flow of information. These formats include:

- the verbal or non-verbal acknowledgements (e.g. confirmation nods, utterances);
- relevant turn-taking (demonstration of knowledge by propagating the topic);
- continued attention – monitoring the partner’s attention indicators (e.g. head direction).

Clark and Brennan (1991) also observed that grounding adapts to different mediums and contexts, but is always present – from the military “yes, sir”, to communication in chat-rooms or over the phone where some of the communicative means removed (e.g. the lack of visual contact introduced “emoticon” text symbols as means of relying emotions in a internet chat session). Between all different formats, the authors concluded that at least one of the following elements needs to be present in order for grounding to be identified by communicating partners: co-presence, visibility, audibility, cotemporality, simultaneity, sequentiality, reviewability and revisability.

The subtlety of the process gave rise to the concept of the *back-channel* (Argyle, 2013). One study identified three main features of the back-channel feedback as that it *i*) responds directly to the content of an utterance of the other participant *ii*) is optional and *iii*) does not require acknowledgement (Ward and Tsukahara, 2000). If the grounding process is the goal, back-channel actions are the active tools by which we accomplish it, without interfering with the main flow of information. Brunner (1979) identified three levels of meaning that a feedback back-channel could have, with the higher level implying and containing the lower ones. These are: *i*) involvement, *ii*) level of understanding, *iii*) actual response.

Even with the above-mentioned mechanisms in place, we are not always certain that we understand each other. An experiment which tried to determine human emotion from a video of facial expressions (Gehrig and Ekenel, 2013) ran into unexpected problems when humans were not able to agree on the observed emotion. Given only short videos (0.3 – 5sec duration, EmotiW dataset (Dhall et al., 2013)), 5 human evaluators achieved agreement on 53% of classifying between the seven basic emotions (anger, disgust, fear, happiness, sadness, surprise, neutral). To the contrary, research on “thin-slicing” (Ambady and Rosenthal, 1992; Ambady et al., 2000) showed that people are good at judging inter-personal consequences from expressive behaviour based on 30 second samples. Our conclusion is that even for human observers, additional contextual information is needed in order to understand their peers, or in the words of Vinciarelli et al. (2008) “... social signals are intrinsically ambiguous and the best way to deal with such problem is to use multiple behavioural cues extracted from multiple modalities”. In real world scenarios, situations can become even more confusing with the presence of non-congruent signals which we perfected into an art of itself (e.g. irony, exaggeration, ridicule).

2.2.1 Digitalization of human interactions

Even though back-channels can be easily identified in a face-to-face conversation (e.g. utterance such as “*m-hm*” in English, “*vraiment?*” in French, or “*aha*” for Serbian speakers), the concept also occurs in digital interactions such as on-line chat-rooms, e-mail, forums, etc. Different formats also caused the information to become less transient and more actionable (McNely, 2009), as Yardi (2006) proposed, for community building, encouraging social interactions, etc.

Digital environments are at the current forefront of exploring new forms of human interactions. The ease of implementation of non-invasive monitoring, discrete and recordable actions (e.g. measurable clicks) and huge audience numbers give us the opportunity to get a different look at various aspects of human activities. For instance, in the educational domain technological solutions enabled us to consider scenarios such as teaching without the teacher (intelligent tutoring systems; Anderson et al. (1984)), or in the opposite direction - teaching to imaginary students (e.g. distance education; Holmberg (2005)). The newest question the technological solutions are focusing on is the scale at which a lecture can be held, and whether it is possible to transmit knowledge with quality to vast numbers of students (Massive Open On-line Courses (MOOCs), Clow (2013); Daniel (2012)).

Majority of human interactions still take place outside of the digital domain, which does not mean they are outside of our scope of interest – rather the opposite. Several overlapping fields emerged focusing on different aspects of social behaviour.

Affective computing (Picard, 2000) focuses on the perception and taking into account emotions as a relevant input in Human-Computer Interaction (HCI). The emphasis on the “human” part of the equation made us re-consider the format of computers. The field of **ubiquitous computing** (Weiser, 1991) emerged twenty-four years ago to accurately predict technologies stepping away from the spotlight and merging with the rest of the household items. The format of these technologies range from wearables (devices such as Apple iWatch, Google Glass, and Fitbit) and mobile devices (smart-phones, tablets, e-readers, note-books) to integrating computers into our living space. Although each prediction gained a substantial life of its own, **ambient computing** remains the most ambitious one, described as “people engaging in the interaction with computers in an implicit and indirect way” (Waibel et al., 2009).

However, making computers pro-actively engage with humans is not an easy task – the complexity of human interactions is an obstacle we have been trying to overcome for almost a quarter of a century. Significant achievements have been made in the domain of smart environments. Smart rooms (Figure 2.1), extensively researched in the CHIL project (Waibell et al., 2011) among others, were envisioned as environments which should “connect people, support human memory, and provide meeting support”. A typical scenario studied in smart rooms is a meeting event of a small number of participants (3-6). We can consider two major components of the smart-room approach, comparable to the two stages of the human grounding process: capture and presentation of data. Current efforts are mostly focused on the first stage – the

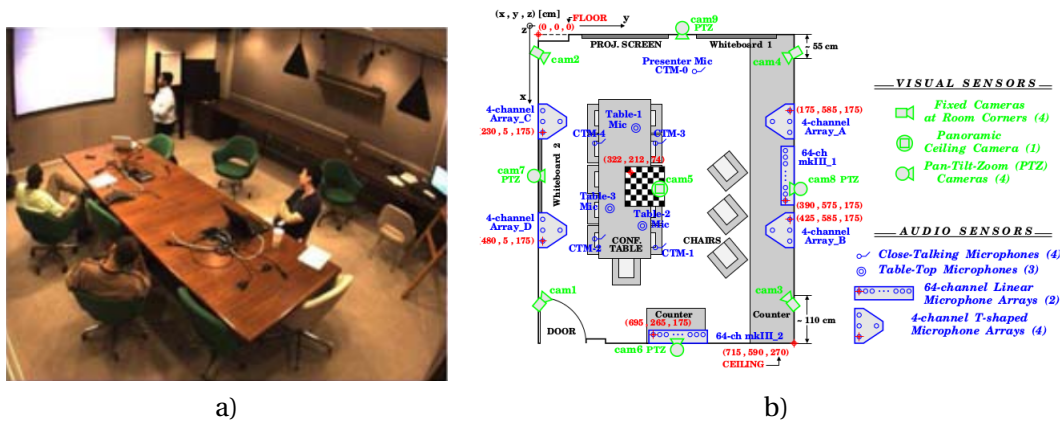


Figure 2.1 – **a)** Camera view and **b)** schematic diagram of the IBM smart room, taken from (Mostefa et al., 2007). The diagram shows the locations of 9 cameras and 152 microphones used for data collection, used for capturing 4-person meetings.

capture of data and the subsequent data analysis, in the service of either aggregating huge volumes of data into meaningful features or shedding light on human behaviour (Stiefelhagen, 2004; Gatica-Perez, 2006; Rienks et al., 2006; Ba and Odobez, 2009; Voit and Stiefelhagen, 2010; Mast et al., 2015).

An interesting aspect of this research is that the focus of analysis moved away from the individual, towards analysing the social interactions between people, known as **social sensing** (or socially-aware computing). The term was first introduced in the work of Pentland (Pentland, 2005; Pentland and Heibeck, 2008; Pentland, 2010). The hypothesis of this research is that quantifying social signals will enable real-time interventions if needed in our communications – situations like regulating the flow of a meeting, connecting unacquainted people, detecting interesting information in the workplace conversations and relationships between family members. One of the principles proposed was the focus on the underlying/subconscious properties of human behaviour with the idea that those features are harder to mimic and will act as more “honest signals” (Pentland and Heibeck, 2008). Similarly to this, Pantic et al. (2011) noted the difference between an *communicative* signal (that is produced in order to convey a particular meaning) and *informative* signal (a signal from which we extract meaning even if it was not intend to convey any). Some of the important manifestations of human signalling that Pentland (2010) listed are: mimicry (reflexive copying of one person by another), activity (as an indicator interest and excitement), influence (affecting another persons behaviour) and consistency (as a marker of expertise), all present in the human interactions without being the direct goal. Newer work on the quantification of synchrony given by Delaherche et al. (2012) demonstrates the broad view of the high-level indicators of the quality of interactions.

A broader overview of the domain has been provided by work of Gatica-Perez (2009). Although limiting itself to small-group interactions, the paper provides four main categories for studying social constructs being: *i)* interaction management, *ii)* internal (cognitive) state of participat-

ing parties, *iii*) personality traits, *iv*) relationships between meeting parties. Apart from the detailed dissection of the technological modelling of different aspects of human interaction provided by the publication, the author also gave insight into the fragmentation of research among different existing fields, which illustrates the complexity of the subject.

As a continuation of social sensing, in an attempt to centralize the research by topic rather than by technology the field of Social Signal Processing (SSP) was formed (Vinciarelli et al., 2008, 2009; Pantic et al., 2011; Vinciarelli et al., 2012). Similarly to aforementioned fields, arguing that machines need to access the vocabulary of social signals in order to fully integrate into human activities, SSP cites non-verbal behaviour as the main source of information needed for the domain. The authors see *social cues*, which they define as “observable changes in facial and body gesture that accompany [human] communication” (illustrated in Figure 2.2), as highly valuable, because with them “humans cannot not communicate” (Vinciarelli et al., 2008) (i.e. they are always present as an integral part of inter-personal communication). The field also made considerable efforts in distinguishing between different social cues, in order to provide a usable systematization for further research. The provided definition of *social signal* declares it as “a communicative or informative signal that, either directly or indirectly, provides information about social facts, that is, about social interactions, social emotions, social evaluations, social attitudes, or social relations” (Pantic et al., 2011). The definition gives us at the same time the difference between four distinct manifestations of social signals, with the common property for all of them (as opposed to “personal” variation of same terms) is the projection of the signal onto another agent – e.g. social evaluation is how we feel about another social actor, as opposed to some random event.

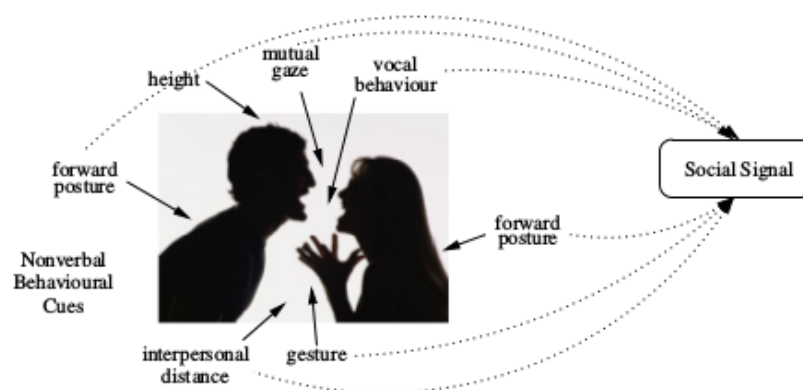


Figure 2.2 – Behavioural cues and social signals, taken from Vinciarelli et al. (2008). The scene, even with the minimal amount of information is clearly interpretable through the cues of gestures and body posture.

2.3 Non-verbal communication

As noted in the previous section, non-verbal communication is crucial for understanding social human beings. The concept of *social intelligence* (Albrecht, 2006) involves among other things the usage of social signals and social behaviours as expressions of ones attitude towards a social situation, and they are manifested through a multiplicity of non-verbal behavioural cues including facial expressions, body postures and gestures among other signals (Vinciarelli et al., 2009).

Major groups of non-verbal behaviour (NVB) that Argyle (1969) analysed in his studies can be classified into:

- general body cues – physical contact/proximity, posture, physical appearance, gestural movement;
- head-oriented features – facial expression, head gestures, direction of the gaze and eye-contact;
- non-verbal aspects of speech – timing, emotional tone, errors, accents.

The same work notes that, even though language is generally associated with cultural belonging, body language is also important for individuals to identify themselves as members of a certain group.

Our non-verbal language relies heavily on the context and usually complements other communication channels. For example, Heylen (2005) found that head movements are notable for co-occurring with speech around 25% of the time. More broadly, Mehrabian (1971) showed that 93% of human affective communication is conveyed through non-verbal means.

In order to map social cues to social behaviours and technologies used for capturing/analysing them, Vinciarelli et al. (2009) provided a systematization shown in Table 2.1. First of all, the list is important to identify the social cues which exist, such as physical appearance and vocal behaviour, but will not be discussed in detail because they lay outside of our research scope. That, however, does not diminish their importance. In case of vocal behaviour, in a salary negotiation scenario (Curhan and Pentland, 2007), analysing the non-verbal speech aspects (activity, engagement, emphasis and mirroring) can predict the outcome (70% classification accuracy) based on the first 5 minutes of the conversation. For detecting high-interest segments of a meeting with the usage of both audio (energy, pitch, speaking rate) and video features (blob detection, hands and head motion, pose eccentricity, rough head orientation), although both modalities had a significant contribution, audio-only features showed better performance than video-only (Gatica-Perez et al., 2005).

	Example Social Behaviours							Tech.		
	emotion	persona.	status	domin.	persuas.	regulation	rapport	speech a.	comp. vis.	biometry
Social cues										
Physical appearance										
height			✓	✓					✓	✓
attractiveness		✓	✓	✓	✓		✓		✓	✓
body shape		✓		✓					✓	✓
Gesture and posture										
hand gestures	✓	✓			✓	✓	✓		✓	✓
posture	✓	✓	✓	✓	✓	✓	✓		✓	✓
walking		✓	✓	✓					✓	✓
Face and eyes behaviour										
facial expressions	✓	✓	✓	✓	✓	✓	✓		✓	✓
gaze behaviour	✓	✓	✓	✓	✓	✓	✓		✓	✓
focus of attention	✓	✓	✓	✓	✓	✓	✓		✓	✓
Vocal behaviour										
prosody	✓	✓		✓	✓		✓	✓		
turn taking	✓	✓	✓	✓		✓	✓	✓		
vocal outbursts	✓	✓		✓	✓	✓	✓	✓		
silence	✓		✓				✓	✓		
Space and environment										
distance	✓	✓	✓		✓		✓		✓	
seating arrangement				✓	✓		✓		✓	

Table 2.1 – Mapping of social cues to social behaviour and technologies used for analysing them. Taken from Vinciarelli et al. (2009)



Figure 2.3 – Equipment for capturing human interactions. Shown in the image are variety of interview cameras, web-cameras, depth cameras, portable eye-tracker, portable EEG, mobile devices (microphone, accelerometer) and input devices.

However, given our scenario of interest, large classroom audiences, video-based technologies are better suited, given that audio input can vary from extremely sparse to extremely complex for analysis. To illustrate with an example from the pedagogical literature, it has been observed that in the traditional classrooms as little as 38.5% of the students engage in active verbal participation, and that 90% of that participation is coming from 1-5 of the most active students (Howard and Henney, 1998). This makes most of our subjects “invisible” to audio analysis.

Large number of features related to human behaviour can be detected by visual means. Capturing devices have decreased in size which allows for their integration into a number of multi-purpose devices such as mobile phones (Figure 2.3). New features were also introduced with the popularisation of depth cameras, such as Kinect (Zhang, 2012) and other *Prime Sense*-based sensors. Manageable incorporation of depth perception introduced improvements to a number of computer-vision tasks such as people tracking, pose estimation, estimating facial landmarks etc.

2.3.1 Posture

Posture has also been shown to be a very informative measure about personal attitude (Richmond et al., 1991), allowing us to predict the affective state of the person. The information is typically acquired by the usage of thin-film pressure pads embodied in chairs (used in (D’Mello and Graesser, 2007; D’Mello et al., 2007; Arroyo et al., 2009)), by observed relative changes in the size of person’s appearance in an image (Campbell, 2009) or by tracking markers positioned on the subject’s body (Dirican, 2014). In the systems where the person was seated, the postural changes at the same time accounted for the distance from the interactive system, which was a good indicator of personal interest (Arroyo et al., 2009; Dirican and Göktürk, 2012). Postures such as “slouching” have been connected to boredom while leaning forward can be connected to “flow” states of the mind (D’Mello et al., 2007; D’Mello and Graesser, 2007; Dirican and Göktürk, 2012).

2.3.2 Hand and body gestures

Numerous hand and body gestures have been known to be indicative of a person’s opinion. An extensive overview of gesture-based interactions given by Karam (2006) showed that hand gestures are the most used type of gestures in human interactions (21% of all gestures). The same work noticed the lack of a global classification of gestures, but pointed that the majority of research is dealing with: manipulations, semaphores, gesticulations, deictic (pointing) and language-based gestures (sign language). Mitra and Acharya (2007) gave an extensive technological overview on methods for gesture analysis from a computer vision standpoint. Some of the most common hand gestures include clenched fist, hand chop, hand scissors, arm folding, self-manipulation etc. (Morris, 1994).

The main problem with conscious gestures is that they are culturally dependant, and because

of that they *i*) lack universality (Mitra and Acharya, 2007) and *ii*) are prone to plagiarism (e.g. usage of body language in acting to convey an emotional state the actor is not experiencing). As the work of Bousmalis et al. (2011) showed, hand gestures are more useful for detecting disagreements in an active discussion, where they act as a supplement to the verbal expression. A similar conclusion was reached by Ba et al. (2009), where the analysis of meetings indicated that there exists a correlation between movement in general with speech in meetings (but not necessarily associated with disagreement between participants). The work of Bousmalis et al. (2013) gave a comprehensive overview of computer-vision (CV) methods which can be used for human action analysis, and interestingly, also showed that the number of usable body cues for detecting disagreement is much larger than those for agreement. For a more complete overview of CV techniques for detecting human gestures, readers are directed to Rautaray and Agrawal (2015).

As shown in Table 2.1, **head-related cues** are extremely beneficial for a number of behavioural analysis. According to the study of Vinciarelli et al. (2009), major groups of cues that can be extracted from the head are:

- **head gestures** – expressive events such as head nodes, shakes, tilts;
- **facial expressions** – display of seven basic feelings (anger, disgust, fear, happiness, sadness, surprise, neutral) as combinations of different atomic facial actions (Ekman and Friesen, 1977);
- **gaze and visual focus of attention (VFoA)** – direction of the head and eyes to capture information from a selected source, as a physical manifestation of attention (the “top-down sensitivity control” mentioned in (Knudsen, 2007)).

Different branches of computer vision have specialized in tasks connected with each of these categories. Given that each task has several overlapping sub-steps (such as detection of head, location of facial landmarks, temporal tracking) and dozens of competing computer vision approaches for solving them, we will not go into a detailed overview of all methods. As its statement of purpose declares, we will rely on SSP literature for an overview of the methods and technical approaches for sampling human behaviour. We dedicate Section 2.4 to visual focus of attention (VFoA) as our primary interest for application in the classroom domain.

2.3.3 Head gestures

Head gestures can be processed as discrete events - in which case classifiers are trained to recognize motion as an explicit gesture. Other researchers like El Kaliouby and Robinson (2005) used both head gestures and facial expressions in order to determine the mental state (they focused mainly on head nodes and head shakes). Similarly, the work of Bousmalis et al. (2011, 2013) used head nodes and shakes to identify people agreeing or disagreeing in conversation. The authors also noticed that head gestures are usually connected with arm

gestures, while Nguyen et al. (2012) observed that using audio cues enabled more precise nod identification (nods are more common when the person is not speaking). Basic head gestures have the properties of being conspicuous and thus easily implementable in affective computing scenarios (Morency et al., 2005).

2.3.4 Facial expressions

Similarly to the general classification of non-verbal cues, we can distinguish between four basic types of facial signals: *i*) static facial signals (structure and biologically-shaped appearance), *ii*) slow facial signals (ageing process) *iii*) artificial signals (items augmenting the appearance, such as cosmetics) and *iv*) rapid facial signals (visually detectable changes in facial appearance). The last category is the main focus of interest for automatic facial analysis for the purpose of social sensing and affective computing, given its capabilities of transmitting emotional information (Pantic and Bartlett, 2007; Gratch and Marsella, 2013).

As with other social cues, facial expressions are prone to falsifications. As noted by Miehle et al. (1973), voluntary facial movements (connected to “posed expressions”) originate in the cortical motor strip, while the more involuntary, emotional facial actions (“spontaneous expressions”) originate in the sub-cortical areas of the brain. The ambiguity between the two is a task which escapes most human observers, but subtle differences might be recognizable by automatic systems. The work of Bartlett et al. (2014) showed that an automated system could distinguish between the simulated and spontaneous pain expressions with an 85% accuracy.

Facial expressions have been shown to be useful in detecting emotional responses to videos, or in the service of affective computing. Work of El Kaliouby and Robinson (2004) managed to reach recognition rate of 87.4% in identifying 6 mental states (agreement, concentrating, disagreement, interested, thinking and unsure), while Jacobs et al. (2009) carried out the work of detecting boredom in individuals watching videos. In high-concentration and high-risk tasks such as driving, detecting emotional distress (Gao et al. (2014), shown in Figure 2.4) or drowsiness (Rimini-Doering et al., 2001) can help prevent accidents. In both cases, the setup dictated that a non-invasive approach was needed, which was implemented by positioning an NIR-camera (near infra-red) in front of the driver.

In education-oriented affective computing, analysing students facial expressions showed to be beneficial when interacting with Intelligent Tutoring Systems (ITS). Whitehill et al. (2008) showed that such systems can predict the self-reported difficulty of exercises by analysing facial features with mean accuracy 42%. In the work of Arroyo et al. (2009), even though the final system was relying on a fusion of inputs from different sensor types, camera-based detection of emotion was the strongest indicator for confidence ($r = 0.72$), excitement ($r = 0.83$) and interest ($r = 0.54$).

Still, the work of Gehrig and Ekenel (2013) shows us that outside of controlled environments and restrictive scenarios, facial analysis can be problematic for humans and machines alike.

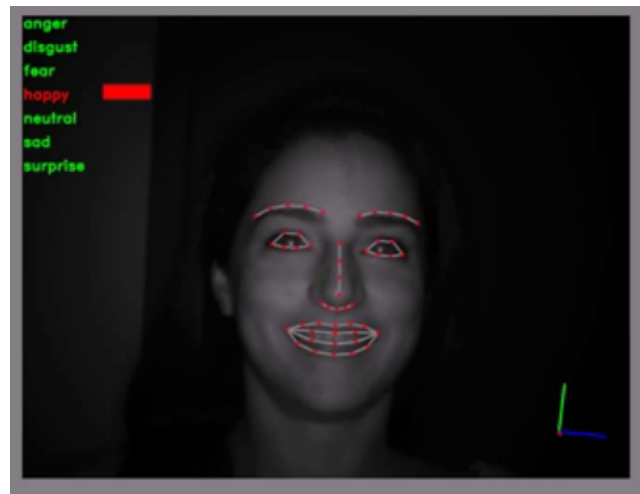


Figure 2.4 – Detecting emotions in drivers. Taken from Gao et al. (2014).

Reported results are usually based on situations where head pose is considered frontal or near-frontal, well lit (good ambient lighting without directional sources of light) and typically adjusted to lighter skin colour (due to easier extraction of facial landmarks with higher contrast of facial features).

2.4 Visual focus of attention (VFoA)

One of the most salient sources of social information is the human gaze. As Chun and Wolfe (2001) put it, “perceiver becomes an active seeker and processor of information, able to intelligently interact with their environment”. Its social usage is developed to the point that we instinctively follow other people’s gaze in order to collect more information about our environment (Birmingham et al., 2008) . This effect is also known as **social gaze**. Still, even a simple indicator such as direction of gaze is not processed in isolation. Even though the eyes are the most meaningful source of information (Emery, 2000), other sources such as the rotation of the head and torso are also taken into account when analysing other people (Langton et al., 2000; Todorović, 2006), and the importance of head and body orientation raises in cases where eyes are not visible (Langton, 2000). Perrett and Emery (1994) named this phenomenon the Direction of Attention Detector (DAD). The significance of cues captured from the eyes is also testified by the fact that studies have found dedicated neurological structures for processing gaze information (Ristic and Kingstone, 2005). Heylen (2005) has shown that perception of other people’s head-signals can be contextual or complementary (back-channel) to the main information channel, usually verbal.

As the most detailed measure of human gaze, eye-tracking has developed as a stand-alone field. With each generation of eye-trackers we progressed from static eye-trackers, IR-based eye-trackers for analysing computer-screen viewing scenarios, to portable eye-trackers (used in our research as detailed in Chapter 6). Eye-tracking established a number of features which

can be extracted from the eyes in close-range such as saccades, pupil dilation etc (Holmqvist et al., 2011). These measurements have been successfully used for marketing purposes (Wedel and Pieters, 2008) and various other tasks.

As informative as eye-tracking is, until now its usage scenarios have been limited by the bulkiness of the equipment needed for acquiring the measure, and the general sensitivity of the measures (due to the eye-trackers relying on the IR reflections on the sclera of the eye, strong natural light can interfere with the measurements; calibration steps are needed on per-person basis). Similar to this, computer vision techniques for acquiring gaze information aimed to handle the problem without dedicated hardware. With the limitation that the subjects need to be seated in front of a computer, with both eyes visible (Asteriadis et al., 2014) reported standard errors of 6.9° .

In the situation where gaze estimation is not possible, either due to technical limitations or to the flexibility needed for a specific usage scenario, scientists have switched to estimating the Visual Focus of Attention (VFoA) based on the head's rotation and position. The assumption for using this measurement as a substitution of the gaze data is that the head will be on average rotated in the direction of the main point of interest. Previous work has provided evidence of this in meeting scenarios (Stiefelhagen and Zhu, 2002; Voit and Stiefelhagen, 2008; Ba and Odobez, 2008b), controlled experiments with animals (Freedman and Sparks, 1997) and invasive experimenting on humans (Guitton and Volle, 1987). Our own contribution to this subject is detailed in Chapter 6.

2.5 Head pose estimation

We will give a short overview of approaches and steps needed for head pose estimation. This is in no way the complete work on the subject. Readers are directed to publications such as Murphy-Chutorian and Trivedi (2009) for a broader overview of the subject.

2.5.1 Problem formulation

In order to discuss existing approaches for head pose formulation, we will start by discussing the **state spaces** used. State space is defined as the set of values that a process can assume. In our case, this reflects not only what ranges of values do we consider for our variables, but also their dimensionality. The abstract notion of dimensionality can usually be connected to fundamental questions about our problem – is head represented by a single floating point value; is the head always observed; how many heads can be observed at the same time? With each new dimension, our prediction system becomes more flexible but also potentially more difficult to train and more calculation-intensive. We will illustrate the formulation choices and how they affect the balance between benefits and difficulties.

Depending on the scenario, our initial branching depends on whether we are dealing with a

single subject (such as the HCI scenarios where the users are sitting in front of a computer terminal (Fanelli et al., 2012; Asteriadis et al., 2014)), multiple subjects (Smith et al., 2008; Cristani et al., 2010) or large crowds (individuals can not be clearly distinguished) (Conigliaro et al., 2013a).

In case of multiple persons, we have a choice of modelling them together as a group, known as the **joint formulation** (Smith et al., 2008; Shitrit et al., 2013) or as a set of individuals modelled in the same way (Ba and Odobez, 2008a). Joint-state formulation comes with additional problems such as varying size of the state-space (in case we allow subjects to leave the observed area) and increase in processing requirements. However it provides means for handling difficult tasks, e.g. identity-preserving tracking under occlusion (Shitrit et al., 2013). In cases where there is a very clear separation between targets, the group of people can be treated as a set of simpler, repeated problems.

For any individual person, head pose estimation can be treated as either a 3- or 6-DoF (degree of freedom) problem, depending on whether we are estimating the 3-dimensional spatial position and rotation, or just the 3D rotations of the head. We will use the annotation (x, y, z) for the spatial coordinates, and (θ, ϕ, χ) for angular rotations (Figure 2.5).

In the case of localization, there is also the question of whether we are localizing the position of the head in the 2D camera-plane location, the 3D world camera-centred location or in a global 3-dimensional coordinate system (Voit et al., 2006; Ba and Odobez, 2008a; Voit and Stiefelhagen, 2010) typically used with multi-camera set-ups. In case of 2D-locations, other possible parameters are the size of the head (Ba and Odobez, 2008a) and ratio between width and height of the bounding box. In scenarios with full-person tracking, the location of the

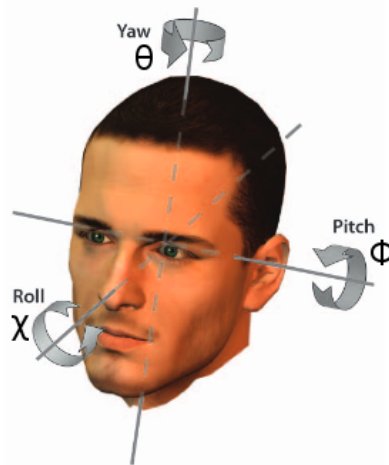


Figure 2.5 – The axes of head rotation are **horizontal rotation** (yaw or pan, θ), **vertical rotation** (pitch, ϕ) and **sideways rotation** (tilt or roll, χ). Image taken from (Murphy-Chutorian and Trivedi, 2009).

body is also used in order to limit the subspace in which the head can be localized (Smith et al., 2008; Cristani et al., 2010).

Various scenarios have taken the liberty of eliminating one or more dimensions in order to simplify the problem or to speed-up the calculations. Other simplifications can be connected to the representations of any of the dimensions. For instance, in Voit and Stiefelhagen (2010) the spatial locations were discretized into voxels (3D pixels), or more frequently, the rotational angles are represented in discrete steps (Ba and Odobez, 2004). Similarly, in scenarios with predefined interactions, we can also substitute the problem of face angles completely with the discrete set of targets (or Areas of Interest – AoI) (Ba and Odobez, 2006; Ba et al., 2009).

2.5.2 Technical approaches

First technical decision to make is the type of input which will be used. Considering only camera inputs, the problem can be formulated as *i*) monocular problem (single camera used) or *ii*) multiple-camera. In addition, different techniques can use alternative sensors such as the Kinect depth-sensor (Fanelli et al., 2012; Mora and Odobez, 2012; Funes Mora et al., 2013), stereo-depth image acquisition or IR-sensor for situations with difficult lightning (Gao et al., 2014).

Head pose estimation is a complex problem that includes several sub-tasks. Two basic steps are *i*) **head detection**, and *ii*) **pose estimation**. Different approaches used can add other intermediate steps such as **facial landmark detection** (Zhu and Ramanan, 2012); **tracking** (Ba and Odobez, 2006); or a separate “VFoA recognizer”-step in case of a set of discrete viewing targets (Ba and Odobez, 2009).

Two main steps (detection and pose estimation) can be processed sequentially (output of the detection step is the input for the pose estimation step), or in parallel (suitable for joint formulation; Osadchy et al. (2007); Smith et al. (2008)). In case of multiple-camera setup, another question is how to merge the information from several sources. This can be treated in various ways, ranging from simple voting for viewpoint with the highest scoring detector (Voit et al., 2006), weighting pose hypothesis from several sources (Voit et al., 2008), to integrating all sources into a joint formulation (Shitrit et al., 2013).

The detection step in itself has been the subject of many publications. In controlled settings, detection can be either completely skipped (with the assumption that there is a single person present in the centre of the image; Wu and Toyama (2000)) or simplified to the level of detection of the biggest skin-coloured blob (Voit et al., 2008). Ever since the influential work of Viola and Jones (2001), head detectors have been increasing in complexity, with the goal of improving robustness and coverage of pose-space.

After (or in parallel to) the detection step, **pose estimation** is handled in a number of approaches. Systematization given by Murphy-Chutorian and Trivedi (2009) classifies the techniques into categories:

- **Appearance template methods** assume appearance similarity between a detected head image with a previously known bank of models, and predicts the angle based on the similarity. The positive side of this approach is easy extension of the knowledge base of the estimator by adding new samples. Two major drawbacks are bad performance with increased number of samples and lack of interpolation options for inputs which are not covered by the knowledge base.
- **Detector arrays** rely on a set of detectors, each trained for a specific angle of the face, with the winning detector determining the pose of the face. A simplification of this would be a combination of front-face/side-face detectors present in modern packages such as OpenCV (Bradski, 2000), and comparable to some of the multi-camera setups (Voit et al., 2006). The good side of the approach is that individual detectors can be implemented in various ways, with the constraint that the individual outputs have to be comparable for the final voting on which detector has the best score.
- **Non-linear regression methods** create a regression mapping between a set of extracted features about the face and the pose angle. Typical examples of the method are SVR (support vector regressor) and multi-layered perceptron (Stiefelhofen et al., 1999; Stiefelhofen, 2004; Osadchy et al., 2007). Different formulations can handle both discrete and continuous pose outputs and in different configurations work with near-field and far-field images.
- **Manifold embedding methods**, similarly to previous formulation, treat pose estimation as a function between a high-dimensional space of head appearance and a low-dimensional space of head poses. Any dimensionality reduction techniques can fall into this category (such as PCA, kPCA) with various pre-processing steps aimed at eliminating the “noise” of irrelevant data in the input.
- **Flexible models**, in which a generic non-rigid model is fitted to the observed image in order to eliminate the personal inter-subject variability, after which the pose can be estimated on the “standardized” representation. Apart from the previously mentioned PBM method (Zhu and Ramanan, 2012) (primarily used as a detector), other examples (and theoretical predecessors of PBM) are Elastic Bunch Graph (Lades et al., 1993), Active Shape Models (ASM) (Cootes et al., 1995) and Active Appearance Model (AAM) (Cootes et al., 2001).
- **Geometric methods** are most similar to human perception. Geometric methods rely on the concept of geometrical symmetry of human faces to observe cues such as nose location shift for determining head orientation. Different sets of facial landmarks are used, typically combinations of nose, eyes and mouth points, with different assumptions of the arrangement (assumed co-planarity etc). Given a reliably located features of the face, the method is very simple and fast.
- **Tracking methods** are significant for incorporating temporal data into estimating the orientation of the face, by accumulating small angular changes. The downside is that

the methods sometimes assume obligatory starting pose, or manual initialization. Techniques employed can vary from tracking facial landmarks, estimating affine transformations for a given facial deformation, optical flow etc. (Wu and Toyama, 2000)

- **Hybrid methods** represent a wide category containing all approaches which employ a mixture of several techniques, such as: clustering and particle filter tracking (Ba and Odobez, 2004, 2008a).

Different techniques have historically emerged as solutions for previously observed difficult situations. Since we already discussed the detection step in the previous section, we will not emphasize on the difference between techniques which need the location of the head as the input for the pose estimation step, although the benefits of approaches which combine the two are obvious.

Physical phenomena such as directional or uneven lightning can pose significant problems for algorithms (Wu and Toyama, 2000), which most proposed methods avoided by assuming good lightning conditions.

Even though the resolution of input images relies mainly on the technical characteristics of the camera, the division between the near-field (large resolution images) and far-field (low-resolution images; e.g. 20x20 pixels image size per subject) problems typically refers to the distance between the capturing sensor and the subject. The distance is reflected in the resolution of the image in a way that far-away targets have smaller resolution than the close-by targets. The algorithms adjusted to far-field situations usually deal with the face in an appearance-based fashion (without trying to interpret the parts of the face), such as appearance template methods or multi-layered perceptron (Stiefelhagen, 2004). With higher resolution images, higher abstractions (and clearer input data) can be achieved with sufficient reliability, and methods such as geometric estimation or flexible models become more usable.

Another source of problems is the variability of appearance, occlusions (obstacles, side-ways rotated faces) and self-occlusions (hand over mouth) which can be handled with flexible models, or approaches with high abstraction such as optical flow tracking.

While some solutions are able to handle all of the parameters at the same time (e.g. manifold embedding), other approaches train separate estimators for each of the needed outputs. An example for the later would be the two separate multi-layered perceptrons developed by Stiefelhagen et al. (1999) – one for estimating pitch, and another for pan of the head. The impact of separating the process into several pipelines should be considered in terms of whether the joint formulation can limit the state-space in a meaningful way and additional computational costs which the separation incurs. In the specific case of horizontal and vertical rotations of the head, the relative independence of the two dimensions allows for such separation.

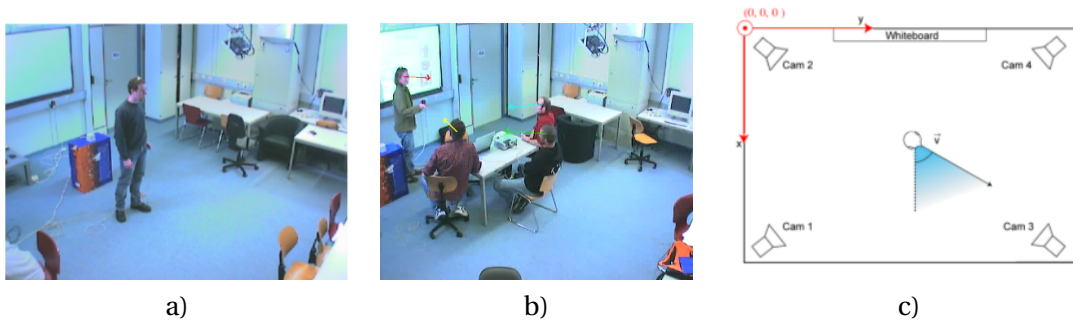


Figure 2.6 – CHIL dataset samples (Waibel et al., 2009). Different scenarios were included from **a**) single person in the room, **b**) to meeting scenarios with **c**) dense coverage of the meeting space from several camera viewpoints. Images taken from (Voit et al., 2008).

2.5.3 VFoA in small group situations

Apart from the “raw” task of head-pose estimation which has been developed by the computer vision community in controlled settings, one of the scenarios in which VFoA has been extensively studied is the “small meeting” scenario. In this setting, gaze and VFoA were used as “important non-verbal communication cue with functions such as establishing relationships (through mutual gaze), regulating the course of interaction, expressing intimacy, and exercising social control” (Ba and Odobez, 2009).

The **CHIL project** (Waibel et al., 2009), which had the objective to “explore and create environments in which computers serve humans who focus on interacting with other humans as opposed to having to attend to and being preoccupied by the machines themselves”, made a large effort by collecting 86 meetings and seminars, with 3-5 people participating. In case when the lecture scenario was considered, the experiment’s goal was mainly to capture the performance of the lecturer, disregarding the actions of the audience. Five dedicated rooms were equipped with a large number of cameras and microphone arrays (depicted in Figure 2.1 and Figure 2.6). The setup was focused on dense coverage of a small space, in which case each person should be clearly separable from others in at least one viewpoint and visible usually in all of them (Figure 2.6c). Scenarios were both spontaneous and acted-out. In addition to audio and video recordings, dialogue transcripts were also provided with the dataset, as well as video annotations, e.g. the location of faces in the camera view.

Similarly, the **AMI dataset** (Carletta et al. (2006); shown in Figure 2.7) collected a dataset of around 100 hours of 4-person meetings. Most scenarios included 4 participant discussions with or without the usage of slides, although moving participants were also included later on. The behaviour of participants was not scripted. Apart from several cameras and microphones, the dataset also included speech transcriptions and slide-change annotations.

In different stages, both projects were concerned with different tasks. In the more controlled settings, the goal was to demonstrate the feasibility of the scenario and creating a dataset for estimating head pose, by placing a single person wearing magnetic motion-sensors in the empty space of the office (shown in Figure 2.6a) (Voit et al., 2008). Similarly, Ba and

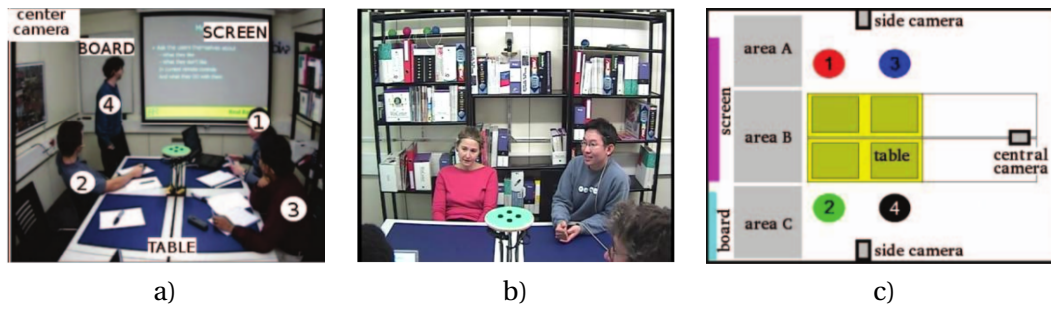


Figure 2.7 – Augmented Multi-party Interaction (AMI) dataset (Carletta et al., 2006). **a)** Overview of the space, **b)** one of the dedicated cameras for two of four participants and **c)** general arrangement of the space. Images taken from (Odobez and Ba, 2007; Ba et al., 2009).

Odobez (2006) also used magnetic tracker for capturing the ground-truth of head pose, but in a full-scale meeting scenario.

In the case of AMI dataset, spatial arrangement played an important role in several aspects. For processing, several mitigating steps could be applied. Because the used camera views allowed for clear geometrical separation of each person (Figure 2.7b), presence of a single head in each part of the video was assumed (Ba and Odobez, 2008b), and no tracking was needed (Voit et al., 2008). Also, the work done on the dataset is characterized by per-location settings and analysis, typically differentiating between the left and right-side person visible in the video.

In the work on meeting analysis, formulation of VFoA was treated both as a classification task, and a continuous measure. In some scenarios, such as Ba and Odobez (2008a), the pan angle was represented as a discretized set of values, and achieved as low as 8.8° errors. Similarly, the previously mentioned approach of Stiefelhagen (2004) with two neural networks reached 52% of correct classifications on 13 categories of horizontal rotation. One of the downsides of the AMI setup was that due to the different positions of persons in videos. Recognizing the difference between recorded subjects in the video, the models for assessing head angles were adjusted for each individual position (Odobez and Ba, 2007). This principle was elaborated in (Ba and Odobez, 2009) with the introduction of unsupervised model adaptation – by batch-processing the videos, best parameters were found for each person and used for re-evaluation of head pose. Same work reported significant variations from one person to another, pan errors ranging from 7° — 30° . For the same reason, Voit et al. (2008) trained separate NNs for each location (left and right person in the video) in order to achieve a mean pan error of 14° .

The other approach discussed, determining the VFoA target from the set of possible targets, was made possible under the assumptions about the activity (people participating in the meeting have a shared, limited number of target choices), and geometrical arrangement of the targets (each of the targets is associated with a distinct head rotation). In evolving experiments, the initial set of 4 targets (3 participants + a table) (Stiefelhagen et al., 1999) or just the participants (Stiefelhagen et al., 2002) was expanded. Ba and Odobez (2008b) included the options of looking at the projector area and being “unfocused”.

Later work (Ba et al., 2009) also considered the effects of slide-changes. The same study formulated the presenter as the moving target located in one of 3 standing zones (diagram showed in Figure 2.7c). The major observation was that parts of the meeting which were not considered crucial in the earlier explorations received high probabilities of being the centre of attention. Table as the target of VFoA was assigned with 30% chance of being looked at at any given moment, while slides attracted attention immediately after changes (50% chance of viewing) which diminished as the time passed from the slide change. Ground-truth for the experiment was acquired by human annotators based on the views from the far-field cameras, which raises some validity concerns.

In the case of dynamic scene analysis done by Voit and Stiefelhagen (2010), the scenario including one moving and three seated persons, with additional passive targets such as the table and the projection area (Figure 2.6b). The dataset included 10 meetings captured with multiple cameras, 7-10 minutes long. A considerable effort was put into formulating the human viewing model. The used formulation relied on volumetric representations of targets, which took into account the occlusions when determining which target was “hit” within the gaze uncertainty region. Even though reported prediction rate is 72.2% of correctly classified targets, the conclusion of the paper was that human field of view was too variable to be accurately modelled without gaze information, and that ambiguous situations such as the presenter standing next to the projection screen could not be reliably handled. Acquiring ground truth was handled in the same way as with the previous studies, using human annotations based on far-field cameras.

2.6 Usage of measurements

Considerable research efforts have been dedicated to the development of methods for automatic observation of human behaviour, some of which were presented in the previous sections. Parallel to data collection, different scenarios driven by the collected data have been developed. Here we will briefly consider several situations divided into single person and group scenarios, building up to our environment of interest – the classroom.

Our emphasis remains on the non-obtrusive measurements (Webb et al., 1966) or indirect collection of data, without disturbing the subjects. Even though the roots of the approach lie in the psychological/sociological domain where they were originally envisioned for manual execution, the automation of data-collection has giving new scopes and applications which would be too difficult for human observers (Mast et al., 2015).

2.6.1 Individuals

With the recent rise of the *quantified self* movement (Wolf, 2010), it is clear that people want to collect data about themselves in order to improve. In our performance-oriented society, data collection has migrated from professional environments to almost every aspect of human life,

be it personal productivity (WakaTime, 2014), physical activities (Fitbit, 2007), usage of public spaces (Williams et al., 2015) or life in general (HeyDay, 2013) .

We have already noted the situations where VFoA and other behavioural measurements can be used as input in HCI, and attention is considered a valuable enough resource that application's response should be tailored to it (Roda and Thomas, 2006). In on-line meeting environments, gaze information was used as input to the system to provide cues about which participant is talking and about what (Vertegaal, 1999), or to what extent was the material of the on-line lesson followed (Sharma et al., 2014). Processing gaze as input in human communication with an interactive system (or robot), two-sided approach has been tested (taking into account the orientation of the head, and location of salient objects) of determining the gaze-referenced object on the table (Yucel and Salah, 2009; Yücel and Salah, 2009). In office environments, social sensing was proposed as the solution for better management of joint space, with the environment adjusting to different communication scenarios between co-workers (Danninger et al., 2005).

With the increasing sophistication of vision-based measures, application domains are expanding to complex subjects such as presentation skill performance and creativity. Echeverría et al. (2014) showed that with pose and gaze data, the presenter's performance can be classified as good/bad with mean correctness of 64.5%, with similar performance achieved by incorporating the cues from audio and presentation slides (Luzardo et al., 2014). Combining multi-modal data streams has proven to be beneficial in experiments with group analysis (Aran and Gatica-Perez, 2010), and we can expect similar benefits in new contexts.

Visual analysis of actions can give us insight into the mind of an expert and a beginner. Work of Worsley and Blikstein (2013) towards analysing the differences between the two groups found observable differences in the hand gestures during an object construction task. Beside the measurable differences (e.g. in hand usage ratio), the insights gathered about the grouping of atomic actions, and resolution of data capture are significant for further exploration. In this concrete study, fine-grained classification of actions did not give the expected, clear separation between experts and beginners, but rather collapsing the groups and considering the context of each atomic action has proven to be much more insightful. The importance of this and previous work is that such novel approaches are helping us answer very difficult questions such as “what makes us good in certain field” and how to judge the performance in open-ended, creative tasks (Worsley and Blikstein, 2014).

With intelligent tutoring systems, various measurements captured can indicate frustration and potential learning difficulties (Whitehill et al., 2008). Given a sufficient number of sensors and financial resources, the emotional assessment can be carried on the entire classroom (Arroyo et al., 2009).

“Group analysis” can easily be mixed with analysing large number of separate individuals. Example of the latter would be an “audience” of a television show which is analysed for marketing purposes (Distante et al., 2014). This includes quick and unobtrusive assessment

of various demographic information (gender, age, identity), modelling in-store or on-line people behaviour, etc. Also, extraction of the mental state from videos of people's faces (El Kaliouby and Robinson, 2004) or gaze behaviour (Soleymani et al., 2012) has proven useful in analysing the impact of video material (e.g. commercials, advertisements). We do not deny the usefulness of these metrics, but the analysis of the *group* in its formulation acknowledges the connections that the individuals have among themselves.

2.6.2 Groups and interactions

The switch from individuals to group increases the complexity of interactions roughly proportionally to the number of individuals in the group. In social sciences, the switch is marked by a transition from psychology to social psychology and sociology, and specifically by considering social interactions (Argyle, 1969) and group dynamics (Forsyth, 2009). Among the myriad of aspects that can characterize a group of people, we will give a brief overview of previously considered from a computational point of view.

Working in pairs serves as the smallest sample for studying human interactions. In case of collaborative work, dual-gaze studies showed that the “chemistry” between pairs can be judged by the synchronization of their gaze (Richardson and Dale, 2005), which was found to influence social intimacy (Cook, 1977). In tasks such as pair-programming, the measure of such synchronization can differentiate between good and bad collaboration (Nüssli, 2011; Jermann and Nüssli, 2012). In other situations, interactional dissynchrony can lead to detection of deception between people (Yu et al., 2013), while observing mimicry can identify pairs which are getting along (Bilakhia et al., 2013).

Human interaction in a number of face-to-face scenarios can be enriched with additional analysis. The questions of agreement and disagreement have been analysed in cooperative situations (discussions; Nguyen et al. (2012)) as well as antagonistic (interviews; Marcos-Ramiro et al. (2013); public debates; Bousmalis et al., 2011, 2013). Even mundane scenarios such as dating (Pentland, 2010) provide a rich source of behaviour cues to decipher.

In small groups (4–6 persons), several aspects differentiate a good collective from a dysfunctional one. Pentland (2010) found that group cohesion was the central predictor of productivity. Gatica-Perez (2006) has categorized different group aspects (ordered from shorter to longer durations) into: *i*) addressing, *ii*) turn-taking, *iii*) interest level and *iv*) trends (such as dominance). Unobtrusive analysis can be directed at each of the social “resolutions”. Automatic estimation of group cohesion can act as a warning indicator (Hung and Gatica-Perez, 2010) or speed-up annotation of interesting parts of a discussion (Campbell, 2008, 2009). Even without higher-level estimation of the quality of cooperation, measuring features such as the amount of verbal participation and showing them as visual feedback to the participants can help the group become more balanced (Bachour, 2010). Similar techniques can identify individual influence and dominance over a situation (Aran and Gatica-Perez, 2010).

As the group size varies, the cues considered change with the “resolution” of the captured data. The loss of a single highly indicative signal can be replaced and approximated from other sources. For instance, even though eye-tracking would provide a very meaningful information, it becomes unavailable in open-space environments. Technologies have been developed to estimate gaze information from head/pose combinations even on moving individuals (Smith et al., 2008). Security and surveillance approaches are looking into integrating social cues in order to better classify suspicious behaviour (Cristani et al., 2010).

Other areas such as crowd analysis are developing far-field analysis of people’s behaviour. In extreme cases, when people can be detected just by their silhouette, motion synchronization has proven to be a good indicator for detecting agreement in groups of spectators (the choral effect; Conigliaro et al. (2013a)). This enabled researchers to classify salient events in the game and which team was responsible for them just by observing the reactions of the audience (Conigliaro et al., 2013c,b).

As Social Signal Processing has identified, in order to improve our algorithms, we need to replace raw feature extraction with domain knowledge (Vinciarelli et al., 2012). Even though brute force mining for statistically significant features can push test-set results upwards, and flexibility of certain formulations can provide considerable coverage of input values, progress achieved with this seems to be at its upper limits.

We already came a long way from “is there a man in the picture?”, over “what is he doing?” and reached “is this a functioning group?”. Questions like “who is a good student?” are beyond our reach, so we focus on the middle ground – how can we support teachers into being better lecturers for their students.

2.7 Classrooms

At different times a teacher needs be a demonstrator, motivator, administrator and evaluator, just to name some of the aspects of the profession. The conflicting nature between some roles can cause misunderstandings between the teacher and the students, as it was noticed in some publications (Hargreaves, 2000; Howard and Henney, 1998; Marks, 2000). It is up to the teacher to restore the balance of a cooperative relationship and maintain it regardless of the students current performance.

Yet, despite potential obstacles, teacher’s impact on student lives is undeniable (Whitcomb et al., 2008; Coe et al., 2014). Teaching can be described as emotionally (Corcoran and Tormey, 2012) and cognitively demanding (Emmer and Stough, 2001). Hattie (2008) highlights the requirement that, for a successful lesson, the teacher needs to be aware of every student of the classroom. With their attention divided between the teaching material, personal presentation and noticing student responses, we see why it is easy to slip into mental short-cuts (Kahneman, 2011), such as attribution errors (Fennema et al., 1990). One of the differences spotted between expert and novice teachers, is their approach to students: while novice teachers focus on

the teaching plan, expert teachers adjust their approach to the comprehension of students (Westerman, 1991). This emphasizes both the importance of teacher becoming a reflective practitioner (Schon, 1984) and the fact that the ability is not one which is acquired easily.

Judging only by non-verbal expressiveness, both experts and novices have equal difficulties of “deciphering” student’s state of mind (Jecker et al., 1965). Observable signals include previously mentioned, such as social gaze (duration of looking at the teacher, frequency and speed of looking away, blinking), head-arm gestures (chin rubbing, mouth movement) and facial expressions (brow raising or furrowing), which illustrates the first needed component – sending of social signals. Understanding students’ reactions depends on a number of circumstances, among other things how long are the student and teacher acquainted (Stader et al., 1990). This illustrates the second component – reception and understanding of social signals. This draws a parallel with previously mentioned levels of meaning of the back-channel that Brunner (1979) identified (involvement, level of understanding, actual response). In order to evaluate the effectiveness of their teaching, teachers are constantly trying to establish a back-channel with a large population of students.

In the literature of teacher training, professional feedback to the teachers has been found to be highly valuable (Hattie, 2008), as a form of deliberate practice (Anders Ericsson, 2008) and a part of a broader knowledge-building cycle (Timperley et al., 2008). However, Range et al. (2013) observed that in order for the intervention to be successful, it needs to be well timed (long delays between action and feedback weakens the association needed for learning), and it largely depends on the experience of the persons giving feedback (Bernstein, 2008).

2.7.1 The shape of the classroom

Traditional classrooms remain the dominant environment for lecturing on all levels of formal education today (Moore, 1989). The geometry of the classroom has been presented as an emotional barrier for more natural interaction (Hargreaves, 2000). Students in the front rows are perceived as “more interested” (Daly and Suite, 1981), and majority of communication is oriented to the T-shaped region with the highest concentration of interaction focused on the front-centre of the classroom (Adams, 1969). This does not only affect the teacher’s perception, but students also adjust to the geometry of the classroom. Students who seek interaction with the teacher will tend to sit in the high-interaction places (Altman and Lett, 1970). It has been also shown that the seating arrangement will act as an amplification of students interactions – making high-verbalizers more active in the high-interaction zone, and making low-verbalizers even less active in the low-interaction zone (the edge regions of the classroom) (Koneya, 1976). The classroom environment greatly affects the perception of the teacher and students, but not always in favour of the learning process.

On the “student-centric” side of research, Daum (1972) found that distance from the teacher also has a significant correlation with the success of students. Finn et al. (2003) found that smaller class sizes (less than 15 students) affect the quality of the lecture in two ways: the

teachers takes less time to manage the learning process, but more importantly the students' interaction between themselves also changed for the better. This seems closely related to students becoming more accustomed to studying in a large group, where individual visibility is questioned and situation makes diffusion of responsibility and social loafing easy (Forsyth, 2009).

2.7.2 Attention in learning

Several attributes have been connected with measuring student's engagement: *attention* as the quality of interaction (Csikszentmihalyi, 2014), *daydreaming* as the periods of extreme dis-engagement (Lindquist and McLean, 2011) or *time-on-task*, a measure more connected with digital environments (Kovanović et al., 2015).

In classrooms, teachers are trained to raise attention with various means and techniques. Breed and Colaiuta (1974) found that basic visual contact can raise the attention of students, and that teachers can be trained into recognizing non-verbal cues of low understanding (Stader et al., 1990). Goldin-Meadow et al. (1992) confirmed that, even with small children, non-verbal language can indicate if the child understands the lesson. Various structuring tools such as mixing activities (Middendorf and Kalish, 1996) have been suggested among various other approaches for maintaining attention (Davis, 2009). Techniques for classroom management, such as classroom orchestration (Dillenbourg and Jermann, 2010; Dillenbourg et al., 2011) have been introduced to help organize the many aspects teacher needs to fulfil in order to successfully manage the pedagogical scenario.

Moore (1989) pointed out that students' attention already is divided between three types of interactions: *i*) learner-content, *ii*) learner-instructor, *iii*) learner-learner. The second of these has the priority over the other two in a lecture, due to its limited availability. But irrespective of position or grades, students have difficulty maintaining the attention during the whole duration of a lecture (Rosengrant et al., 2012). Even if it is not clearly quantified after how much time students lose attention, proposed values vary between 10 minutes (Wilson and Korn, 2007) and 20 minutes (Middendorf and Kalish, 1996), far shorter than the average duration of class period. Population-wise, it is reported that between 33% (Geerligs, 1994) and 54% (Cameron and Giuntoli, 1972) of students are not attentive during the class, with better attention percentage associated with the smaller class-size (Finn et al., 2003). And as we can intuitively guess, daydreaming episodes during classes have been negatively associated with student performance (Lindquist and McLean, 2011).

Assessing attention during the class has been the focus of several previous studies. The most wide spread technical aid for this purpose are the clickers (Caldwell, 2007), dedicated devices which can serve as a tool for sampling students opinions and cultivating a peer-instruction atmosphere (Mazur, 2009). Although not directly sampling attention, their purpose of identifying misconceptions during the lesson greatly aids the reflective side of teaching.

Other approaches such as “live interest meter” (Rivera-Pelayo et al., 2013), are using the omnipresence of mobile devices to receive feedback about the quality of the lecture, as perceived by the students. A similar approach was developed by Holzer et al. (2014), with mobile devices as a platform for real-time anonymous feedback. The downside of both approaches is that they are relying on the students will to provide feedback, and are introducing additional administrative devices and protocols unrelated to the actual subject of learning, thus increasing the complexity of the learning process for students and teachers alike.

2.8 Conclusion

In order to tap into the students state of mind, we have made a case that attention is beneficial for the learning process. That a person’s state of mind is manifested in non-verbal behaviour, and that the vocabulary of non-verbal language is complex. Technology for coping with this complexity is progressing on various fronts by analysing human interactions, but it needs to be adjusted to different scenarios in order to meaningfully model human behaviour.

In the next chapter we will lay the theoretical base for the systematization of this research, and emphasize what we think is important when observing interactions in the classroom.

3 Channels and signals of classroom interactions: an informational view

BY studying subject's interactions with their surrounding, we gain knowledge about the subject's intentions, properties and habits. Focusing on this indirect way of studying human behaviour Webb et al. (1966) formulated the concept of **unobtrusive measures**. The authors went into great length showing the potential of studying traces of human interactions. The main rational behind this approach is that, although more complex for implementation, it steps around subject's mental traps and has the potential to be more honest.

But unlike counting the number of book check-outs from a library, or measuring the erosion on the carpet of a museum, classroom interactions do not leave a physical trace. The way this is typically handled in education is by formalizing the communication by either:

- distributing written tests, which are primarily used for documenting the grading process and, not that often, as a teaching device,
- migrating parts of the educational process to on-line learning platforms, which in turn provide richer interaction and detailed tracking of student activities,
- introducing dedicated devices into the classroom which are used for feedback purposes (clickers, mobile device apps, tablets, etc.),
- scripting interaction, in which case the protocol guarantees participation (Dillenbourg, 2002).

However, all of these techniques introduce an organizational overhead, and/or significantly alter the original interaction.

For this reason, our research is focused on capturing human interaction in the classroom, inspired by the concept of unobtrusive measurements. The modern touch on already existing methods is the ability to condense raw data (hours of video material) into a more meaningful set of metrics which would be both faster to produce and analyse.

Chapter 3. Channels and signals of classroom interactions: an informational view

In order to provide a systematic overview of our efforts, we borrow a number of concepts from information theory, all of them centred on the idea of the *communication channel* as the flow of information from one person to another. To illustrate the final goal of the guided learning experience we also introduce the *entropy* of the classroom environment in two variations – *i*) in the information theory sense of the signal's property, and *ii*) in the thermodynamic sense as the measure of chaos or order in the system.

Classroom research is already laying on the intersection between psychology, sociology, and pedagogy. We can question whether we need to add information theory as another ingredient to the mix. The decision is not driven by the desire to make things even more complicated, but to give a set of metaphors which will help us decode human interactions and guide our research by emphasizing the principles behind the observations presented in the following chapters.

Even though classroom environment allows for great flexibility in the format of the lesson, our focus is on lecture-style interactions – teacher presenting in front of a large audience (more than 20 persons). This was chosen as one of the most frequent scenarios in present-day education. As mentioned in the previous chapter, group-work and working with smaller number of students have already been discussed in a number of other publications and contexts.

The reason why the lecture scenario is interesting is that it is inherently sociofugal – it focuses the attention of a large number of individuals on a single presenter (teacher) who has to infer the state of mind of that audience while carrying out the primary task of lecturing (properties of reflective practitioner; Schon (1984)). Neither of these comes naturally to us, even when we are facing such a problem in the professional capacity. Thus, we aim to aid teachers by strengthen their relationship with students, and not replace.

3.1 Elements of classroom communication

We focus on a set of key elements needed to explain classroom interactions. We see classroom interactions as **signals** transmitted over **channels** between **transmitter** (source) and **receiver** (sink) in order to pass **messages** between each other. We recognize some key aspects of each of the concepts, most importantly – that acknowledgement of a received message is a natural mechanism present in human nature. This concept is known as “grounding” in social sciences (Clark and Brennan, 1991), carried over the “back-channel” (Yardi, 2006; Heylen, 2005). However, we will use the term **synchronization** between source and sink to generalize the concept. Our approach uses the fact that synchronization is a publicly observable event, in order to verify whether the messages are indeed reaching their destination (i.e. whether the students are listening to the teacher).

3.1.1 Sources of messages

The definition of an emitter is a machine or a device that emits messages (a source of messages). Messages in our case are units of information, which could further be discretized into Shannon's *bits* (Shannon, 2001) if we were dealing solely with technical devices. A classroom imposes a set of contextual roles and rules, which allow us to further develop our list of emitters or sources of messages:

- we have two distinct roles in the classroom – teacher and students;
- a number of learning resources is present – a blackboard, a projector, books, notes; each of them with different reach (some are meant for private and some for public viewing).

The classroom in “lecturing mode” allows us to make an assumption that there is a single main activity taking place, and that participants have a varying interest to participate in it. With this in mind, a student has the option of being “tuned in” to either:

- an internal process (thoughts),
- a personal resource such as a book, notebook, mobile phone, magazine etc.
- another student – either actively (chatting, discussing the lecture) or passively (e.g. seeing other people listening to the lecture),
- a public resource such as a blackboard or a projector,
- the teacher.

Our list is a bit more detailed than the one proposed by Moore (1989) because we needed to further split the previously proposed categories “student” and “content” in order to differentiate between different behaviours.

A source of messages does not need to be intentional or to be emitting messages by design. Equally good examples of messages directed at a student would be a definition told aloud, a sudden silence of the teacher, or a simple yawning of a fellow student. Each of them carries clear information, although not equal relevance.

Depending on whether the source of messages can adjust in an unscripted manner we differentiate between:

- **reactive sources** – teacher and other students in the classroom,
- **passive sources** – books, projector and other objects,

with all meaningful interactions containing at least one reactive source.

3.1.2 Signals

Signal represents the encoding of the message. Among many properties of signals, we identify two important aspects – modality of the signal and its base. Example of signal classification based on these criteria is shown in Figure 3.1.

		Modality	
		Audio	Video
Base	Verbal	spoken word	text
	Non-verbal	music, noise, audio signals	images, body language

Figure 3.1 – Signal classification based on modality and structural base.

Modality is more frequently used and well established in the technical literature. Its importance in our context is shaped by human perception – while audio signals are omni-directional, video signals are directional. That means that in case of a video-based signal we have a physical constraint (gaze contact) needed for the signal's propagation.

The second dimension (base) is proposed because of its connection to the signal's structure. Signals always carry semantics, even if it is trivial, but they do not always possess a syntax, as noted by Vinciarelli et al. (2012). While verbal-based signals are defined with a set of rules needed for their understanding (grammar, syntax), non-verbal signals usually do not have a clearly defined vocabulary. This raises the uncertainty connected with the signal's decoding.

In information theory this can be further formalized as the signal's **entropy** – the probability that a single message conveys new information, with rare signals conveying new knowledge having large entropy, and frequent and predictable signals having low entropy (i.e. low informational value). Entropy in the communication between two machines is quantifiable given that both sides are using a common, previously established grammar and entropy is defined by the probability of occurrence for a given message.

We define entropy in human communication in order to identify one of the sources of misunderstanding between conversation participants. Entropy's main component lies in our ability to interpret received signals, which is in turn dictated by our cultural background, knowledge, mood, etc. It also changes with our familiarity with the source of the signal, and so signals sent between long-time friends can be simpler but with much higher entropy than signals received from strangers or people with different cultural backgrounds.

3.1.3 Channels

Channel is a single-modality connection between two end-points which is used to convey a signal. Depending on its role in the channel, an end point can either be a source (sender of the signal) or sink (receiver). Except for technology-aided channels, human communication

traditionally takes place over a common medium (“*aether*”). Because of the medium’s non-restrictive properties, we consider communication to be public in nature.

A person can dedicate several channels to the same source (e.g. listening and looking at the same person), or can divide them between different sources (listening to the teacher while looking at the slides). As an illustration, we can draw the parallel between the concepts of *bandwidth* and *working memory*. Two high-bandwidth channels (e.g. teacher’s voice and textual information in the course-book) will cause the receiver to ignore one of them due to the mechanism of *competitive selection* (see Section 2.1).

Note that despite characterizing human communication as being public, the channels described in this sections have the *one-to-one* property of interaction. This goes against the multicast or broadcast concepts which we typically associate with a lecture. Our interpretation is that the lecturer still tries to communicate with individuals in the audience – which is illustrated by how teachers monitor their students and by the mental effort needed for maintaining the contact evenly across the classroom.

For a successful communication between two people, one of the channels is usually dedicated to synchronization (or “grounding”) between participants. For this reason we see communication as a two-way connection over one or more channels in which

- the source sends the signals and receives confirmation,
- the sink sends confirmations and receives the signals.

With our goal of confirming the propagation of signals in the classroom, we emphasize that each side of the channel is a sender of signals which we can tap into to observe the informational flow. Sociological literature also recognizes the two-way communication between individuals and the “secondary or background complement to an existing front-channel”, also known as the *back-channel* (Argyle, 2013).

3.1.4 Synchronization

From a conceptual perspective, synchronization has to have a small footprint because it carries a small amount of information and needs to be frequent in order for the conversation to continue. This makes it perfect content for signals transmitted by body language, which is not typically relied on as the main communication channel (sign language excluded). Another important feature of using body-language for synchronization is that, by utilizing one of the less-used signals, we are making sure that we will not have overlaps between confirmation messages of the back-channel and main stream of information coming through the front-channel. If visual contact is not maintained, the synchronization can be carried over by verbal means, where it keeps the same properties – low complexity, short and relatively frequent confirmations (“*a-ha*”). The disproportion between the bandwidth of the “front-channel” and “back-channel” makes the communication asymmetrical in nature.

Different message sources in the classroom require different types of synchronization, which we can split into two basic categories:

- **direct synchronization** (synchronization with the source) is the synchronization which originated from human interaction, and is partially present in the previously mentioned concept of grounding. We can generalize it as the receiver's reaction that follows changes in the signal, or an active adjustment for better reception of the signal. A typical example would be a student nodding to the teacher's explanation or turning her head in response to a pointing gesture;
- **indirect synchronization** (synchronization with other receptors) is available only when we are discussing a group of receptors. The assumption is that if the receptors are "tuned-in" to the same source, they should react in a similar way. The difference from the previous type of synchronization is that while direct synchronization is observed in the relationship between the source and a receptor, indirect synchronization is observable when comparing several receptors to each other. An example would be the entire audience of a cinema screaming simultaneously while watching a horror movie, or the whole classroom laughing at a teacher's joke.

Synchronization signals in the classroom, as anywhere else, are prone to misinterpretation. This is an effect caused by the relationship between participants, where the student population is susceptible to manipulate the signals in order to appear more knowledgeable. This is doable because the teacher's back-channel has the spotlight effect, and the individual student needs to send confirmations only for brief periods of eye-to-eye contact. For that reason, the informational entropy of the individual synchronization signals is very low (a nod confirms very little).

On a bigger scale, indirect synchronization can be compared to a measure of order (entropy) in a thermodynamic system. It is important to notice the difference between the two types of synchronization – while direct synchronization serves as the confirmation of receiving a signal, indirect synchronization illustrates that there is a pattern in the system, but without clear indication to what it relates to. A well-managed classroom has low entropy (an orderly behaviour) because of the teacher's influence, but a group of sleeping students also has low entropy because they agree that the lecture is not worth paying attention to. In order to judge the feedback from an audience as positive, a combination of both synchronizations needs to be present.

3.2 Spatial dimension

Another important aspect of the classroom is its spatial arrangement. It was already noted how space shapes our perceptions of communication (Sommer, 1959), and classroom environments are no exception to that. In existing studies, classroom space was normalized and

divided into standard regions (front, middle, back; left, centre, right - (Adams, 1969)). Since the perception of the teacher changes significantly depending on how far students are from the front, we decided against this kind of standardization and based our formalization on the proxemic theory of zones (Hall, 1969).

Proxemic theory deals with the very general concept of inter-personal space between individuals across cultural differences and activities. The theory defined four specific zones: *i*) intimate, *ii*) personal, *iii*) social, and *iv*) public; based on the social usage of space in human activities. Keeping in mind the relatively static nature of students' positions in the classroom, we defined four specific zones, depicted in Figure 3.2:

- **teacher's space** – the zone in front of the classroom, between the blackboard and the first row of students.
- **immediate neighbour** – which models “personal space”. It covers the person to the immediate left or right of the student, with whom the student shares the desk-space and leg-space. This is partially dictated by the dimensions of the student desks which are often made for two persons per desk;
- **visible neighbourhood** – represents roughly the zone two rows in front of the student and ± 2 people wide. This represents the “social zone” in proxemics (identified as spanning from 1.2m - 3m). The zone practically models the people who would be intentionally or unintentionally observed by the student who is following the material on the slides or looking towards the teacher;
- **non-visible students** – students who are either too far to the side or behind the individual and can not be seen without intentional action.

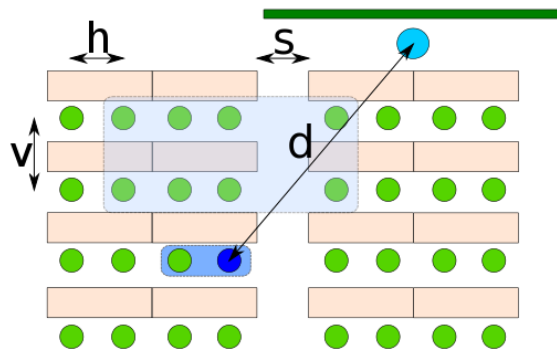


Figure 3.2 – Organization of classroom zones and units of measurements, top of the image represents the front of the classroom. v, h - vertical and horizontal spacing between students is 1 unit of distance (uod). s - between-row spacing, 1 uod. d - distance between the professor (center-front of the classroom) and the analysed student. Light-blue zone represents the visible students for student at the location 3rd row, 4th seat. Darker-blue rectangle around the student represents his immediate neighbourhood.

Except for the generalized concept of “teacher’s space”, the other categories are concerned on the perception of audience members by other audience members. This also puts the emphasis on the interactions that happen in the audience, where we see the main complexity of classroom interactions. We refer to this as a “student-centric” view of the classroom – as we try to judge the classroom events not through the actions of the teacher, but through the reactions of the students.

3.3 Interpreting the classroom

Very much like the unobtrusive measurements which measure “irrelevant” traces to extract meaningful information, we are focusing on the information which is present, but observable only through the right “filters”. With the proposed set of metaphors, we can say that in our research we are focusing on low-complexity, low-entropy signals which are *i)* sociologically and psychologically encoded in our behaviour, *ii)* hard to capture when observing a large number of individuals. Low complexity of signals allows them to be accessible to automatic detection, and low entropy explains why the cues have not been broadly used until now – the difficulty of the task did not permit to use them without technological aids.

By focusing on a very generic concept such as the communication channel, we are unconcerned with the content being transmitted. That remains the teacher’s responsibility. This gives us considerable flexibility needed for any technology aimed at classroom usage. We have defined our approach so that we can answer a formal question – is can we capture evidence of informational flow between endpoints; which is at the same time a very simple one – is everybody with me?

Our measurements focus on the:

- inter-personal agreement, by indirect synchronization of the audience members,
- changes in personal behaviour caused by reception of the signal,
- behaviour coordination between the teacher and the students as a type of direct synchronization.

Our end goal is not to build a detailed interpretation of a single person’s behaviour (increasing the entropy of an individual signal), but to allow the teacher to perceive the “thermodynamic” entropy of the whole classroom as a system. We hope that by making the classroom entropy explicit, we can allow teachers to better coordinate their efforts for conveying knowledge to students.

4 Recorded attention and classroom behaviour samples

Parts of this chapter have been published in Raca and Dillenbourg (2013), Raca and Dillenbourg (2014), Raca et al. (2014) and Raca et al. (2015).

IN this chapter we will present our methodology for capturing the classroom experience and the two different formats used for sampling attention of the students. We will also present conclusions reached from analysing the data in the questionnaires, which will give us some insights about the properties of the classroom. We will briefly compare them to the knowledge present in the pedagogical literature. The findings from this chapter will serve as assumptions about students' behaviour used in the following chapters.

We will take this opportunity to expand on the previous analysis (Raca and Dillenbourg, 2014) by using all of the collected data, and to answer additional questions which occurred during discussions with other researchers.

4.1 Capturing devices

The modern approach to capturing human behaviour “in the wild” can be frowned upon for the lack of control the experimenters have of the conditions, counterposed to the idea of capturing the natural behaviour, which can be changed considerably in the laboratory settings. The benefits of studying classroom lectures are that it is in itself already a well structured and controlled activity. With this, our goal was set to capture as much as we can about the student's behaviour during the lecture.

First constraint of the research was that we did not acquire a dedicated room for our recording sessions. The experiments were conducted in the lecture rooms of the EPFL campus. For this reason, recording equipment needed to be portable and the set-up time for the experiment needed to be relatively short (in practice, time needed was approximately 15 minutes). Main components of the setup included:

- a set of cameras; throughout the experiment 3 types of cameras were used, the preferred

Chapter 4. Recorded attention and classroom behaviour samples

model being the consumer-level web-camera Logitech Pro c910 (4 pieces), but also the Sony MSH-PM5 interview cameras (2 pieces) and the Panasonic Lumix DMC-GX1 (2 pieces) with optical zoom for far seated regions of the classroom. Video material was captured either in full-HD (1080p) or half-HD (720p) quality;

- 4 laptops for capturing the web-camera recording; separate laptops were used in order to avoid frame drops or video corruption due to parallel disk usage;
- numerous support items such as USB extension cords, flexible camera tripods, and two poles for positioning the cameras (Manfrotto static pole).

Although the camera setups were observable, their structure was thin and did not represent a visual obstacle for the students. We take care in positioning the systems in the least obtrusive location. Due to the format, our collection devices were naturally observed by the students in the audience. In order to justify the scenario as ecologically valid, we sampled students opinions in a set of interviews presented in Section 4.6, which were generally favourable.

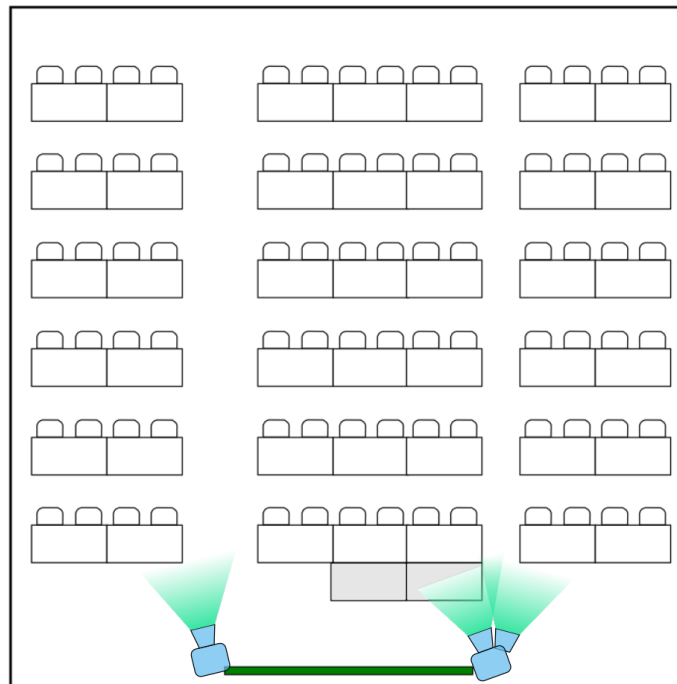
Two main classrooms used in the research (classrooms A and B) had the capacity of 84 and 90 students. Although of similar sizes, the shape of classroom B was pentagonal (layouts shown in Figure 4.2). In order to make the scenarios comparable, the second beamer was turned off during the recordings, which gave us approximately the same viewing angle and teacher zone size for both environments. An example of deployed equipment is shown in Figure 4.1.

Teacher area in both cases was denoted as the area surrounding the blackboard which doubled as the projection area (Figure 4.1). The zone width was approximately 4 meters spanning typically from the teacher's desk on one side, to the opposite side of the projection. Although teachers were free to move in the classroom depth direction, that was not typically observed, and the arrangement of classroom B had no rows between the benches which would allow it.

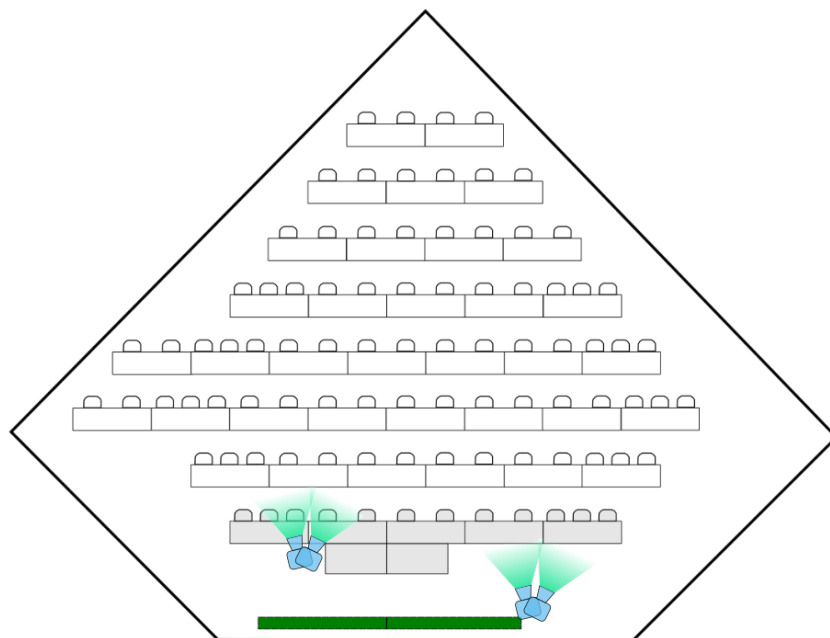
Although majority of the resources was dedicated to capturing the audience reactions, in 4 recording sessions teacher's motion was recorded with a dedicated camera positioned in the back of the classroom. The camera was aimed at capturing the whole front of the classroom, in order not to limit the teacher's motion to the virtual cage we designate as the "teacher's area". Because of this wide-area capturing goal, the trade-off was that the resolution of teacher's personal features was typically very low, and processing was reduced to silhouette tracking (discussed in Chapter 7).



Figure 4.1 – Equipment deployed during the experiment in classroom B.



a)



b)

Figure 4.2 – Layouts of classrooms **a)** A and **a)** B. Camera positions and orientations are designated with bright-blue boxes on the sides of the lecture's area (as used in the head-detection recordings). Gray tables represent the seats which were not used during the lecture.

All classes relied mainly on slide projection accompanied by teacher's narration and interactions with the audience. Although the blackboards were available to teachers in both cases, they were used only for occasional notes. Classes did not contain instances of group-work or scheduled grading examination. No restrictions were imposed on the interaction between the teacher and the students.

As much as the initial part of the research was focused on capturing student motion, the second part was mainly concerned with capturing the facial orientation. The differences between the setup was mainly in the camera positioning, where the first two samples were filmed to provide for better view of the students at the cost of bigger angles, the second part tried to position the cameras very close to the projection area (viewed as the most probable target of student gaze). Due to unforeseen difficulties, each setup contained a problematic element for the other – when capturing motion, the heads were constantly at a large yaw angle, preventing us from capturing the direction reliably; and when capturing head orientation, the oscillations of the poles caused by the teacher's motion combined with the teacher figure blocking the students in the first couple of rows prevented reliable extraction of motion measurements.

4.1.1 Post processing

The whole dataset amounted to 42 videos approximately 45 minutes long captured in 9 sessions. The videos were synchronized and cropped in order to speed-up the automatic video analysis steps. We annotated a number of properties captured in the video:

- **classroom events** – we annotated the visible changes in slides such as slide transitions and animations, as well as the interactions between the teacher and the students (question-asking and answer-giving periods). In case of the attention questionnaires filled out during the class (in-class), we marked the beginning and end of the questionnaire fill-out period;
- **audience information** – each student was assigned a unique identifier and marked with a rectangular region which he or she occupied in the camera view; we also annotated the location of the face in the first frame with the intention of using it as initialization for tracking algorithms, but the practice was dropped as being unnecessary for the techniques used. In the gaze-perception experiments, each student was labelled with the zero-angle (horizontal orientation angle between the student and the centre of the projection area).

Due to the objective nature of the properties, a single coder was used. All data was captured in XML format which was used for all further processing steps.

4.2 Sample statistics

We base our results on analysis of 4 populations of students, described in Table 4.1. The data collection was carried out in the autumn of 2013 (populations 1, 2) and 2014 (populations 3, 4) in the regular university lectures. All student groups were in the bachelor/master programs of EPFL who agreed to participate in the experiments without a compensation. In the lecture before the first recording the experiment was presented to the students as the “capturing of student’s classroom perception” without specifying the exact goal of the recordings. Written informed consent forms were collected. Students were instructed to behave naturally, that they can decline to participate at any time, and that the data collected would be kept anonymous and will not influence their academic achievement in any way. Students who declined participation were attending the class normally, with no obligations of filling out the questionnaires, and their annotated regions were disregarded in further processing.

Students population was gender-mixed, with the female population consisting on average around 32.27% of the population. In the cases of multiple recordings, the attendance had a steady base of 63.38% of all students enrolled, with minor oscillations.

The teachers were two experienced lecturers teaching on subjects from technical sciences in case of Population 1, 3 and social sciences for Population 2, 4. The lectures were given in different times of the day - one being in the morning (1, 3), the other in the late afternoon (2, 4), and in different rooms (each lecturer remained in the same room for the duration of their course).

Cls	No.s.	Cam(s/t/e)	Size	Attend (μ, σ^2)	In/Pst	F-rat	Shape	Intr	Goal
1	1	2/0/0	18	18 (0)	1/0	22.2%	4x5	-	Move
2	1	2/0/0	38	38 (0)	1/0	36.8%	6x10	-	Move
3	4	4/1/1	43	27.5 (6.6)	3/1	34.5%	8x12*	-	Gaze
4	3	5/1/1	62	39.3 (1.1)	1/2	35.6%	6x10	10	Gaze

*Table 4.1 – Basic information about analysed classes. The table shows the number of class/population; number of recorded sessions; number of cameras used (for students, teacher and eye-tracker); size of the population; mean attendance (and variance); number of in-class/post-class questionnaires used; percentage of female population; shape of the classroom (rows/seats in row); number of interviews conducted; the goal of the data collection. * - Number of seats in this classroom varied in every row, mean value displayed.*

In populations 3 and 4 multiple recording sessions were conducted during the semester. This allowed us to treat the data collected as a longitudinal study. For this purpose, each student from the class register was assigned an unique identifier which was kept consistent throughout the semester.

4.3 Capturing changes in attention

Capturing the state of the audience during a lecture balanced between intrusiveness and objectivity of the capturing method. Invasive approaches such as equipping students with eye-trackers (Rosengrant et al., 2012) were not scalable to the classes of 30+ students. We developed two types of questionnaires which we used inter-changeable on two student populations. Questionnaires were developed with the generous input from Roland Tormey and Jessica Dehler, who helped both to formulate the questions and the format of the questionnaire.

We did not use previously developed systems for evaluating public speaking, such as PSCR - Public Speaking Competence Rubric (Schreiber et al., 2012). Even though such solutions are well adjusted for evaluation of the presenter, the main focus of the research is to evaluate the state of the audience as a reaction to the efforts of the speaker.

We also conducted a pilot-study using the clickers (Caldwell, 2007) instead of the paper-based questionnaire. In our experience, the clickers were much harder to setup and more disruptive to the class (students took longer to respond, and were more confused with several questions presented). Even with the prior explanation of the usage, students were confused when several questions were asked in a single fill-out period without visual guidance from the clicker devices (questions were presented separately, on the projector, but the device lacks clear identification which question is being answered). Also, each sampling time required more involvement from the teacher, who needed to incorporate the questions into the lecture presentation. The approach was abandoned after the initial attempt.

The main difference between two used questionnaire formats was the time when the student was expected to fill-out the data. Our first attempt was aimed at simulating strobe-sampling of student attention (similar to attempts of Lindquist and McLean (2011)). We will refer to this format as **in-class** questionnaire. In the second format, the paper was filled-out at the end of teaching period, and will be referred to as **post-class** questionnaire. Questionnaire designs are shown in Figure 4.3.

All values collected with the questionnaires are shown in Table 4.2. Over the period of two years, the format of the questionnaire evolved, with some questions taken out in the attempt to minimize the impact of data-collection on the lectures. This is one of the reasons for the uneven number of samples shown in Table 4.2. Second reason is that students refused to answer some of the questions (typically - classroom attention was usually left out, because students refused to evaluate their peers). The four measures listed as “perception” (attention, classroom attention, teacher’s energy and material importance) were present in all versions of the questionnaire.

With teachers’ cooperation, we also conducted the post-class knowledge test related only to the content shown during that lecture. The questions were prepared by the teachers and projected in front of the classroom, while the students filled out the answers on the back of the questionnaire sheets. The tests were not previously announced, were not collected as

4.3. Capturing changes in attention

Figure 4.3 – The **a) in-class** and **b) post-class** questionnaire used during our experiments.

4.3.1 In-class questionnaires

The procedure included distributing the questionnaires at the beginning of the class. At four equally-spaced moments during the lecture a sound signal was given, at which point the students filled out the appropriate questionnaire section (one of four), indicating their attention, class attention, etc of the previous 10 minutes. The filling out procedure was aimed to take less than one minute, including the time needed for the students to find the paper and set it back away.

Chapter 4. Recorded attention and classroom behaviour samples

Question	Format	No. sessions	No. samples
Perceptions (in- and post-class)			
Personal attention	Likert scale (1-10)	9	1075
Classroom attention	Likert scale (1-10)	9	1025
Teacher energy	Likert scale (1-10)	9	1062
Material importance	Likert scale (1-10)	9	1058
Pre-class questions (only post-class)			
Prior interest	Likert scale (1-10)	3	388
Prior knowledge	Likert scale (1-10)	3	388
Post-class questions (only post-class)			
Test of knowledge	Open-ended questions	3	344
Activities (only in-class)			
Listening	Check-box	2	56
Taking notes	Check-box	2	56
Repeating key ideas	Check-box	2	56
Distracted thoughts	Check-box	2	56
Interacting with others	Check-box	2	56
Using laptop/phone	Check-box	2	56

Table 4.2 – Parameters collected with the questionnaires, with the number of samples and classes captured. Brackets beside the parameter group name indicate the questionnaire format which was used for that group.

with learning (listening, taking notes and repeating key ideas) and the three common activities negatively associated with learning (distracting thoughts, interacting with others without permission and using laptop for tasks unrelated to learning). The students were instructed to indicate yes/no answers for all activities they performed in the previous time segment. The question were later removed in the attempt to minimize the disruption of the classroom.

Overall statistic of the periods captured by in-class questionnaire shows that the average time length of the period was 10.6 minutes, and the average break length used for filling out the questionnaire was 44 seconds. This questionnaire format was used in 6 recordings.

Even though the interruptions were kept to a minimum, this approach to sampling attention was considered intrusive on the learning experience. With the concern that the approach might bias the students in the positive direction (mentally “waking up” the students and making them conscientious about their learning) the post-class approach was developed.

4.3.2 Post-class questionnaires

Post-class questionnaire (shown in Figure 4.3b) has restructured in opinion sampling into 2 blocks: pre-class and post-class, leaving the lecture time intact.

The pre-class block was used as a simplified pre-test, recording students interest and knowl-

edge about the class content (with Likert scales, ranging from 1-10). We did not explicitly ask questions related to the lecture to avoid priming the students.

The post-class block was capturing the same 4 student's perceptions used in the in-class format, with the difference in the presentation of questions. Each measure was entered as a graph with horizontal axis representing the time of the class, and vertical axis representing the Likert scale 1-10 for student's response. The goal was to allow students to be more expressive and freely show if they felt they had a drop of attention in a certain moment. Vertical dimension was also marked with colour gradient to give an orientation of high/medium/low segments. In order to give the students temporal orientation, instead of minutes, the horizontal axis was marked with slides from the lecture (constant speed of presentation was assumed). Even though additional guide-lines were given and the usage of the graph visualizations was explained, most students chose to put "ticks" on the designated vertical time-marks without describing their attention in between the periods.

Because of the temporal distancing, the major concern with this format was the problem of assessing the attention state which occurred 30 minutes in the past. For the same reason, the activity questions were removed, with the assumption that the answers would either become global or difficult to associate with a specific point in time at the end of the class.

In the analysis, the reported levels of attention were associated with 4 time periods of equal length (average duration of period is 11.7 minutes), making them comparable to the information captured with the in-class format. Post-class questionnaire was used in 3 sessions.

4.4 Data in the questionnaires

Given the two questionnaire formats, and the potential objections noted in the previous section for each of them, our first task was to compare the answers between the two in order to continue using both as comparable sources of information in our studies. After resolving this question, we list observations made from data analysis of the questionnaires. Even though this is not the final goal of our research in terms of novelty, the knowledge is further used for forming and eliminating hypothesis about our video-based measurements.

4.4.1 Differences between in-class and post-class format

The main interest of the research remains in the four student perceptions of: **i)** attention, **ii)** classroom attention (how the student perceived the attention of other students), **iii)** teacher's energy and **iv)** material importance. We formed two hypothesis:

- the in-class questionnaire will have a higher mean attention, under the assumptions that the interruptions are waking up the students, making them more attentive;
- the recency effect (perceiving events which occurred more recently as more prominent),

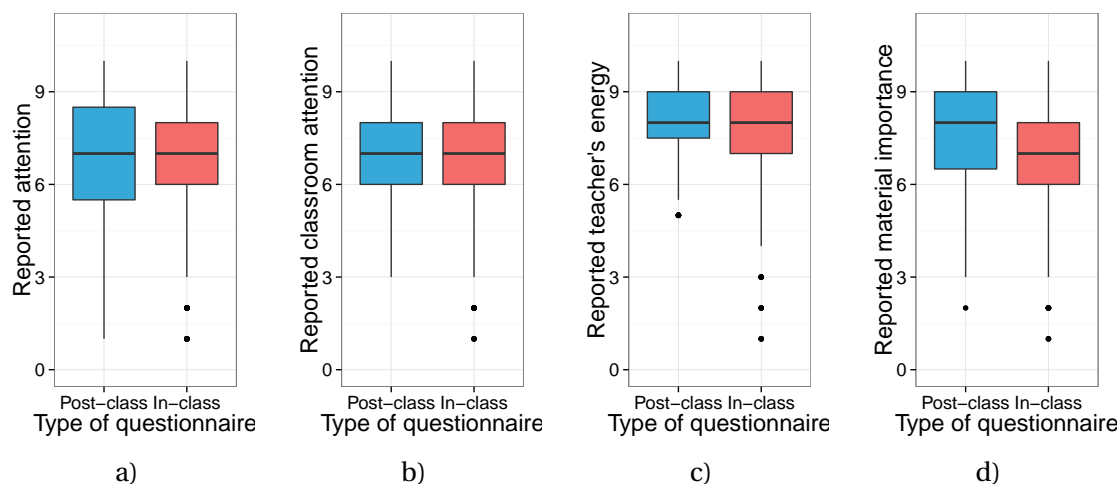


Figure 4.4 – Measurements compared across different questionnaire types. In case of a) attention and b) classroom attention no statistically significant differences were observed. In cases of c) teacher energy and d) material importance, ANOVA showed significant differences between questionnaires (in both cases $p < 0.001$).

will bias the last periods in post-class sampling as more important/attentive.

The comparison of aggregated measures for each of the four perceptions showed that there were no significant differences for in cases of attention (Fig4.4a) and classroom attention (Fig4.4b). Both the teacher's energy $F(1, 1060) = 31.98$ ($p < 0.001$) (Fig4.4c), and material importance $F(1, 1056) = 31.77$ ($p < 0.001$) (Fig.4.4d) had significant differences shown by ANOVA test.

To determine the source of differences in the case of two measurements where it was observed, we compared each of the periods separately. For teacher's energy (not visualized), the difference was observed in all periods except the 1st, with recorded values declining more towards the end of the class (not observed in the in-class format). For the material importance (shown in Figure 4.5a) the post-class was consistently higher scored than the in-class questionnaire.

In case of attention (Figure 4.5b), we observed that in the 2nd period the attention was assessed higher in the post-class format, and variance of data was higher in the post-class condition, but not significantly different. With this, the first hypothesis can be rejected, as the means of both conditions showed no consistent trend.

The second hypothesis was not observed in any of the measures. Material importance in both cases seemed to oscillate very little over time, which leads to the conclusion that students judged the value of material in general, rather than on the material being immediately presented (even though some parts of the lectures contained examples and others lecture definitions, the difference was not recorded in the questionnaires).

We will continue by using personal attention and classroom attention data collected with the in-class and post-class questionnaires interchangeable in our analysis.

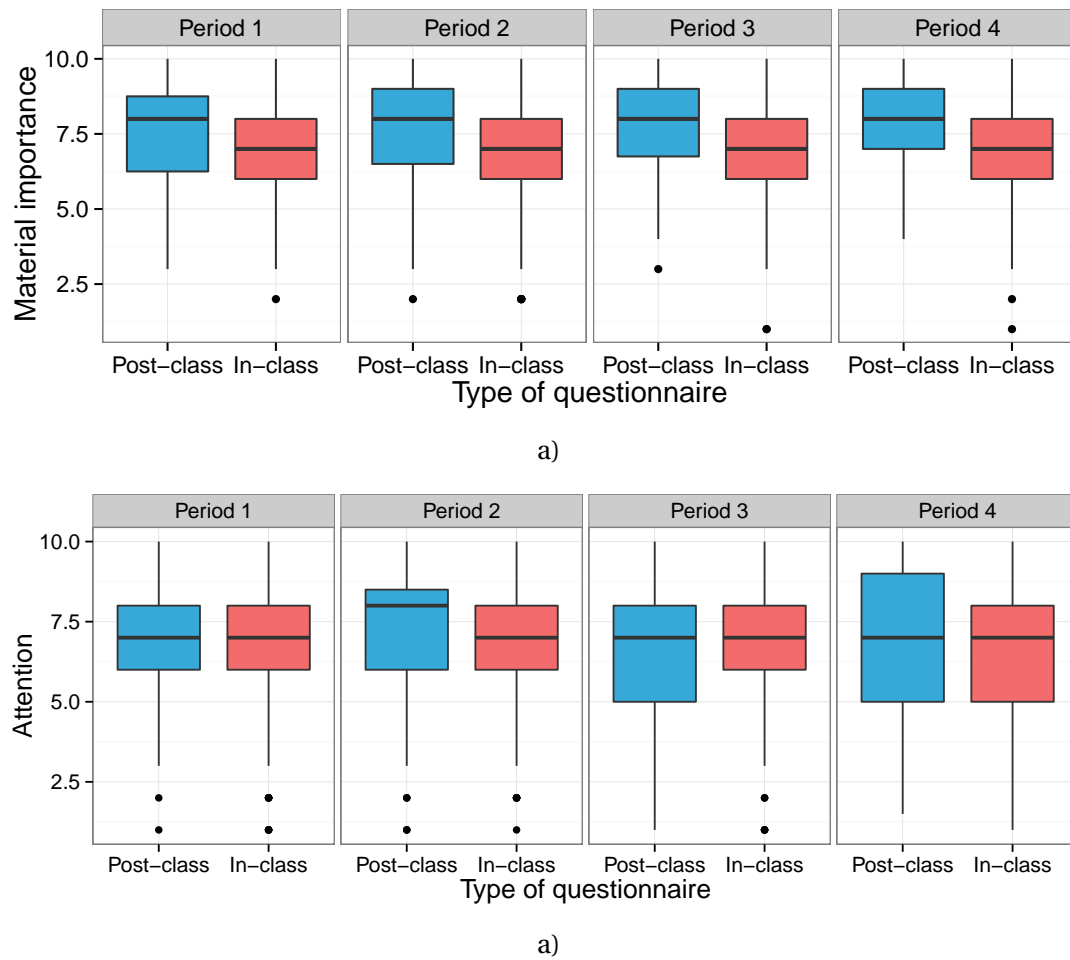


Figure 4.5 – Comparing **a)** material importance and **b)** attention captured over the four periods with the in-class and post-class questionnaire format.

4.4.2 Results captured in the questionnaire

The collected questionnaire data was used primarily as the basis for further analysis of the collected video material. Nevertheless we report the condensed findings to depict the general situation in the classrooms, properties of individual measures and potential interplay between different values.

4.4.2.1 Attention

Reported levels of attention in all populations were high, with the mean value 6.79 ($\sigma = 2.07$) on the reporting scale 1-10 (Figure 4.6). There was no significant difference in student's mean attention between different class populations (visualized in Figure 4.7).

There was a significant difference in reported attention between genders ($p < 0.01$), with

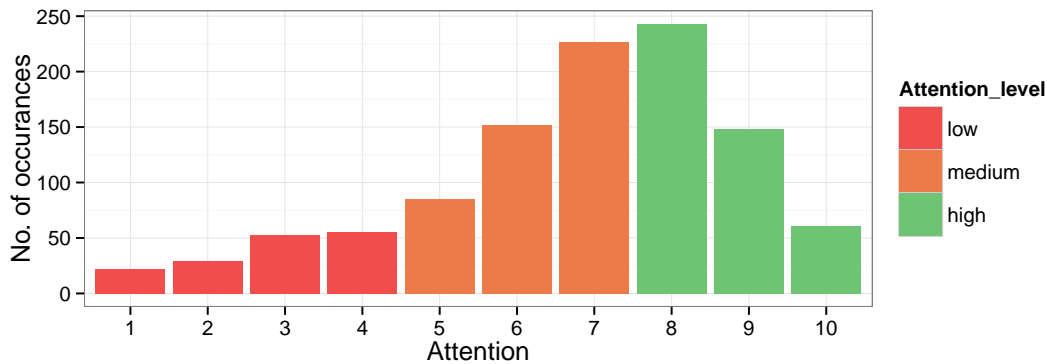


Figure 4.6 – Aggregated attention samples of all students in the study ($\mu = 6.79$, $\sigma = 2.07$). Colours designate the attention level labels associated with each reported level (Number of samples: low (159 samples, 14.79%), medium (464 samples, 43.16%) and high (454 samples, 42.23%))

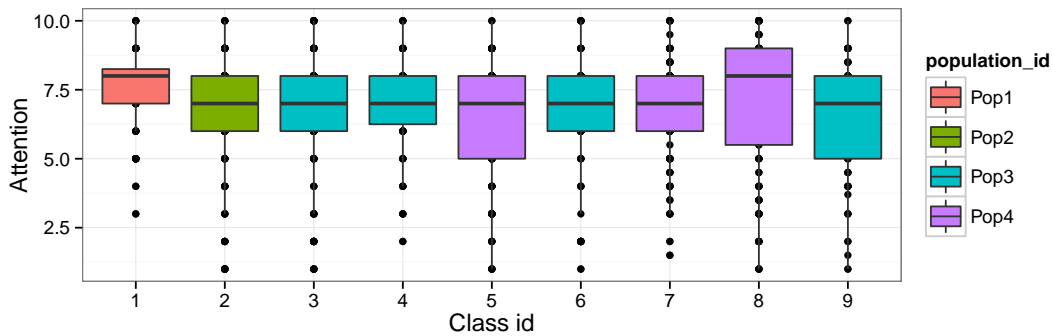


Figure 4.7 – Attentions of all classes, with different student populations visualized in different colours. Bars are presented in the order in which the data was captured. Populations 1, 4 and populations 2, 3 are associated with the same lecturer.

female students having higher attention ($\mu = 7.16$) than males ($\mu = 6.6$). We confirmed that the attention was influenced by the identity of the person $F(1, 1057) = 4.142$ ($p = 0.04$), class $F(8, 1057) = 2.955$ ($p = 0.002$) and the interaction between the two $F(8, 1057) = 3.636$ ($p < 0.001$) meaning that the attention level is not a global property of a person needed to be assessed only once, and that attention will vary from lecture to lecture. Attention per class and person had the average variance of 2.46 points.

Based on the observed distribution of attention, for the purpose of training predictive models we labelled the attention with three labels: *low* (levels 1-4, 14.79% of the samples), *medium* (levels 5-7, 43.16% of the samples) and *high* (levels 8-10, 42.23% of the samples) displayed as the differently coloured bars in Figure 4.6. The split was aimed at mimicking student's interpretation of the attention scale in context of the class.

Our next interest was to test the influence of student's location and time of class on attention. Our analysis of attention over different sampling periods of a single lecture showed no significant trend (Figure 4.8a). Given that our mean sampling period was around 10 minutes, we can

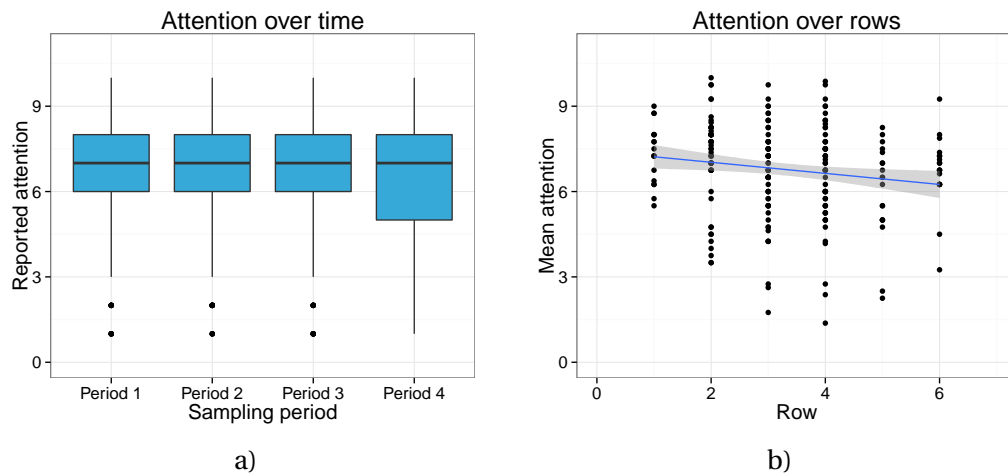


Figure 4.8 – Changes of attention over a) the 4 sampled times (1075 samples); b) the distance from the teacher (225 samples, $b = -0.19$, $r^2 = 0.02$, $p = 0.01$).

report that we found no support for the attention limit of 10 minutes proposed by Wilson and Korn (2007). It can be argued that our in-class sampling was maintaining the attention on the same limit, but the similar trend (no declination) was also recorded in the post-class format of the questionnaire.

Student's row number was negatively correlated with mean attention $r(223) = -0.19$ ($p = 0.01$) (Figure 4.8b). The lateral position (seat number) alone was not significantly correlated. Given that the students had the free choice of seating positions we should consider the possibility that the choice of location was motivated by the student's mood. This is discussed in Section 4.5.

Even though we saw no significant correlation between attention and class period in the feature analysis, we recorded attention levels for Markov property. Transition matrix between 10-level attention is shown in Figure 4.9a. More informative transitions probabilities between 3 labelled attention levels shown in Figure 4.9b. The trend of transitioning to the neighbouring state visible as the strong diagonal in Fig.4.9a enforced even more the tendency of remaining in the same attention level after the binning of values into 3 categories.

4.4.2.2 Other perception measures

Given the relative stability of attention during lecture, we tested the relationship between attention and pre-class measurements – prior knowledge and prior interest. In both cases, we found strong correlation with mean attention, with prior interest showing bigger influence $r(95) = 0.48$ ($p < .0001$) (Figure 4.10a), than prior knowledge $r(95) = 0.26$ ($p = 0.005$) (Figure 4.10b).

Even though we do not try to prove a direct influence of attention on acquired knowledge, our

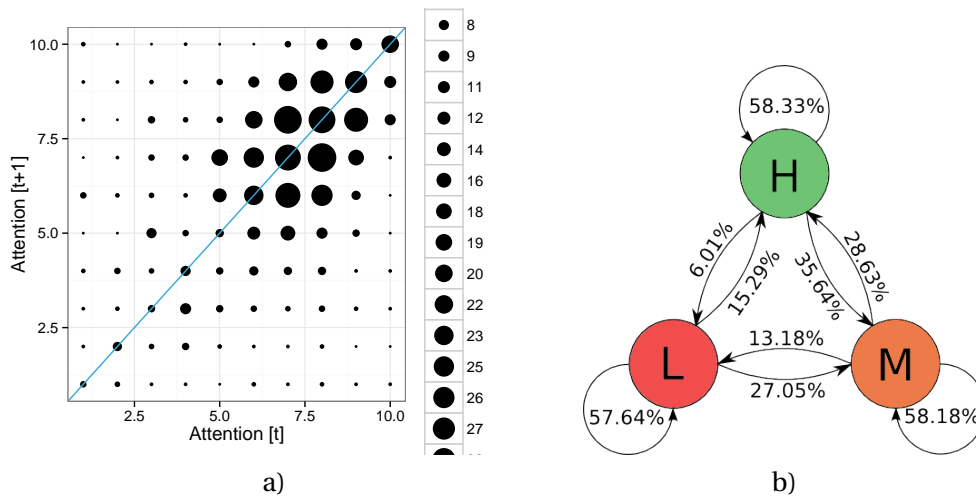


Figure 4.9 – Attention transition values for a) individual levels displayed, increase of attention is represented in the upper triangle of the matrix, while the blue line represent maintaining the same level of attention; b) transitions between discrete levels of attention (low/medium/high) declared in the Section 4.4.2.1.

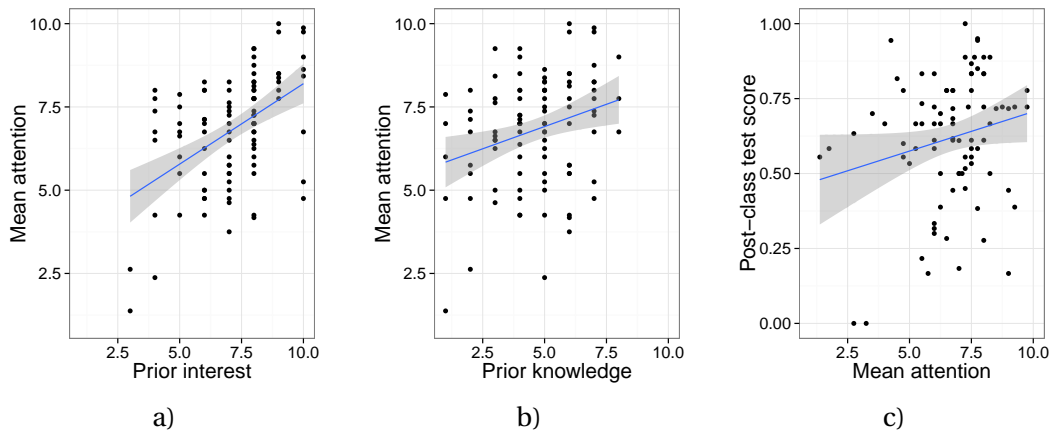


Figure 4.10 – Relationship between a) prior interest and mean reported attention (sample size = 97, $b = 0.48$, $r^2 = 0.22$, $p < 0.01$), b) prior knowledge and mean reported attention (sample size = 97, $b = 0.26$, $r^2 = 0.06$, $p < 0.01$), and c) mean reported attention and results in the post score (sample size = 86, $b = 0.02$, $r^2 = 0.03$, $p = 0.056$).

assumption is that attention is highly beneficial for learning. For that reason we tested the relationship between the mean attention and the post-class test score. We found marginally insignificant positive correlation between the two with a very low influence $r(84) = 0.02$ ($p = 0.056$), displayed in Figure 4.10c.

We tested the relationship between attention and other measurements with the initial hypothesis that there are not going to be any correlations with measures of teacher's energy and material importance, and potentially a weak correlation with perceived attention of the classroom. Our hypothesis was rejected, with all three properties showing strong direct correlation with student's personal attention: perceived classroom attention $r(1026) = 0.43$ ($p < .0001$), teacher's energy $r(1056) = 0.25$ ($p < .0001$), and material importance $r(1060) = 0.41$ ($p < .0001$). All correlations are shown in Figure 4.11.

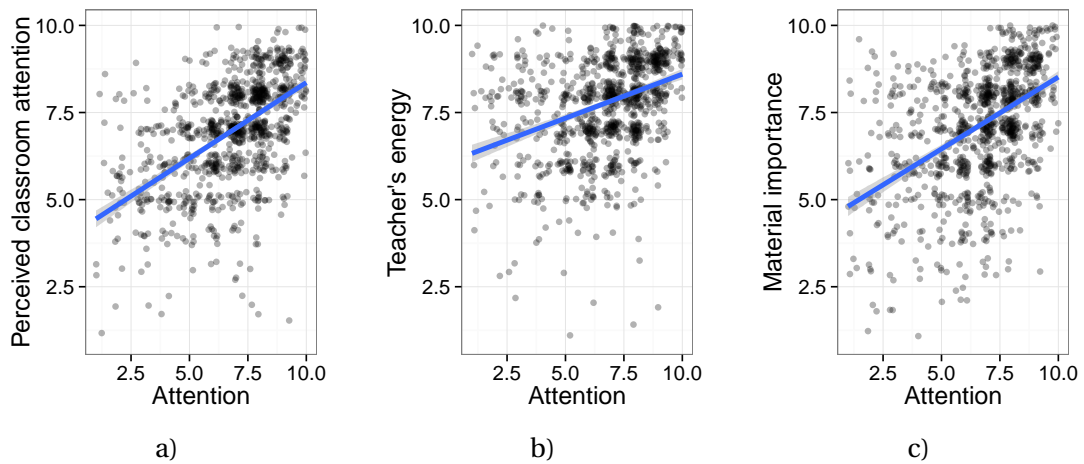


Figure 4.11 – Relationship between attention and a) classroom attention $r(1026) = 0.43$ ($p < .0001$), b) teacher's energy $r(1056) = 0.25$ ($p < .0001$), and c) material importance $r(1060) = 0.41$ ($p < .0001$). To display the density of many overlapping points, Gaussian noise ($\sigma = 0.2$) was added to the location of the points after the linear model was fitted and individual points were made semi-transparent.

4.4.2.3 Student Activities

We relied on self-reporting to capture the information about student's activities for each class period in Populations 1 and 2. Reports show that learning-related activities (*listening to lecture*, *taking notes* and *repeating ideas*) were naturally more represented on higher attention levels. Figure 4.12 shows distribution of activities per attention level. It is also interesting to note that the off-task activities (*distracting thoughts*, *talking to others*) were reported on all except the maximum level of attention.

It's interesting to note that the observable activities in both “directions” *taking notes* and *talking to others* were reported on almost all levels. We did not conduct further studies whether the frequency of those actions in videos was indicative of attention level.

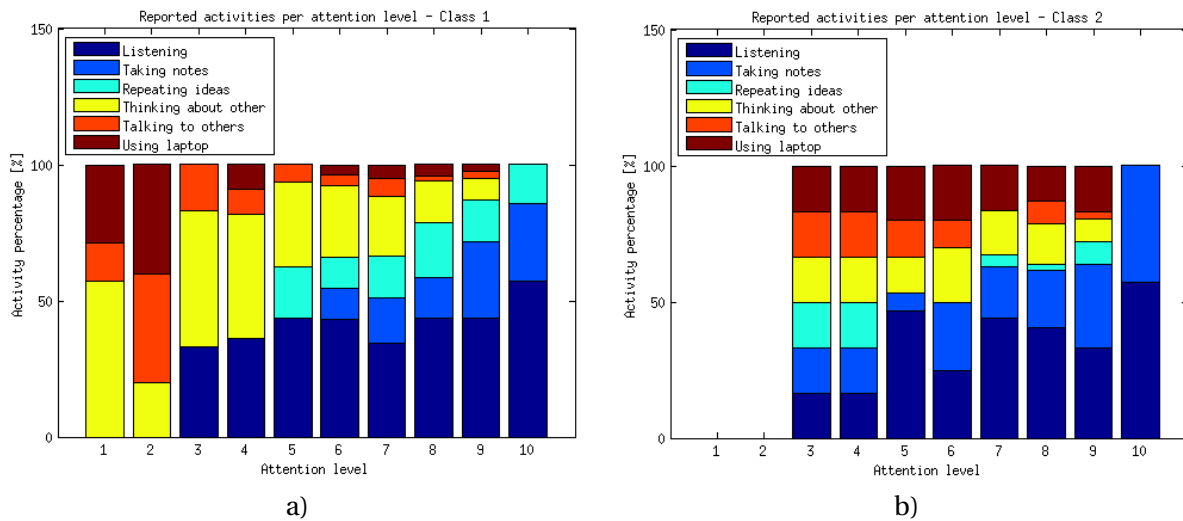


Figure 4.12 – Percentage of reports of each activity per attention level in a) Population 1 and b) Population 2. Number of reported instances was normalized by the total number of instances on that attention level to produce the percentages.

4.5 Mobility of students

With the previous observations that the distance from the front of the classroom was negatively correlated with attention, we tried to test how much did the students change their seating preferences in general. As noted before, students had free choice of seating position in all experiments. Previous studies already observed that high-verbalizers tend to choose high-interaction places as their permanent location (Altman and Lett, 1970). Our hypothesis is that students typically remain seated at the same (habitual) location.

We used the data from the longitudinal part of our experiments (Populations 3 and 4). The number of times specific student participated in our experiment is given in Table 4.3. For our conclusions we used any student who was observed more than once. Every pair of locations which the student occupied was used to calculate the change in terms of row, seat and total distance relocation. Relocation distance was mapped to a Cartesian coordinate system (as explained in Section 3.2) in which both the seat number change and a row number change were used as 1 distance unit changes.

No. of participations	Frequency	Used in study
1	13	No
2	32	Yes
3	33	Yes
4	13	Yes
Total used	78 pers. (215 obs.)	

Table 4.3 – Number of times student showed-up in class during the experiment.

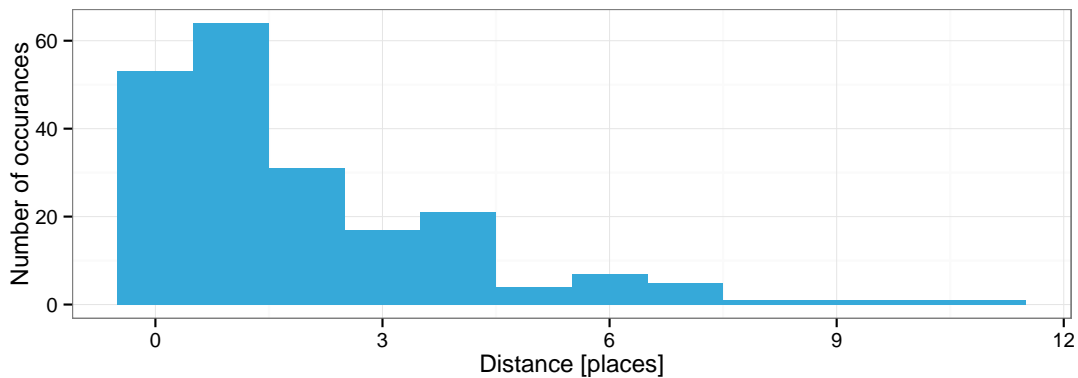


Figure 4.13 – Histogram of student location changes.

The results of observations are shown in Table 4.4. The main relocation distance of 1.91 shows that the students don't typically move from their usual place of sitting, and decomposition of the value between rows and seats shows that switching seats is more likely than switching rows. To better visualize this, Figure 4.13 shows the distribution of overall distances of seating changes captured in our sample.

Also, during the interviews (data presented in the following section) students declared that they do not have a fixed location (80%), but that they were sitting on their preferred location (70%). Our conclusion is that students build their preference in the long term, as previous studies have shown, and do not change their location often. With this, we confirm our initial hypothesis that the students have a habitual place of sitting. We do not perceive student relocations as an influence on lowered attention levels associated with the back of the classroom.

Dimension tested	Seat	Row	Distance
Values [$\mu(\sigma^2)$]	1.91 (6.07)	0.49 (0.65)	2.13 (6.06)

Table 4.4 – Results of mobility of students.

4.6 Interviews with the students

In order to complement our quantitative data with qualitative research of students' perceptions of the classroom, we carried out a series of interviews with the students from Population 4. Participation in the interview was voluntary and monetary rewarded. Due to an unfortunate scheduling (the lecture ended at 7pm), the interviews were carried out the day after the recording in the first available time slot.

We carried out 10 interviews (4 female participants, 6 male participants). Each interview was approximately 30 minutes long, carried out in private settings. Interviews were recorded with the permission of participants and later transcribed and analysed for answers. This allowed the answers to be open-ended and students were encouraged to explain themselves. Although no student openly refused to answer any of the questions, some answers were too vague to

Chapter 4. Recorded attention and classroom behaviour samples

extract a conclusion afterwards. The interviews were anonymous, data collected being marked with the identifier number that was assigned to the person in all recordings.

Interviews touched on the following topics (roughly in the order presented):

- **personal and social context** – how student was feeling on the day of the lecture; how well do they know other people in the class;
- **perception of the lecture** – positive and negative opinions about the lecture (presentation wise), keywords associated with the lecturer (attention was paid to the modality in which the lecturer was described); do they feel that the lecturer is paying attention to them (visibility of audience members);
- **learning practices** – what was the focus of student's attention during the lecture; habits of taking notes; how often do they interact with the teacher during the lecture (asking questions related to the lecture, giving answers etc.);
- **distracting elements** – description of the surroundings (to what extent was the person aware of other students); how often do they have off-topic discussions during the lecture;
- **experiment perception** – how disruptive was the experiment; was the experience unpleasant and if so, why;
- **seating customs** – how did the student choose the seating location and how often do they change the location.

For understanding student's perception of the lecture, we paid attention to the modality students used to describe the lecture elements. The findings (shown in Figure 4.14) show that the audio signal was mentioned more often than the visual. Between different visual targets, slides were referenced more often than the lecturer.

Similar trend was observed when questioning what were the distinctive marks of the presenter. The answer was not guided in order to capture anything that stuck with the students. The overall positive experience in the classroom resulted in a large number of attributes generic to any good lecturer. The division between the visual and auditory characteristics, admittedly influenced by the personality of the specific professor (all students were from the same population), showed that the motion of the teacher made a significant impact on the student's perception as 6 out of 10 participants mentioned it. Vocal attributes were present in the least amount, but were usually focused on the clarity of speech.

In terms of interaction with their surroundings, people who reported talking had lower attention (3 persons, $\mu_{talking} = 6.0$), compared to the persons who did not talk (6 persons, $\mu_{silent} = 7.5$). All "talkers" reported talking to a single neighbour.

People with higher mean attention described fewer neighbours (estimated $b = -0.64$), but the correlation was not significant. Almost all subjects (9 out of 10) remembered to describe their

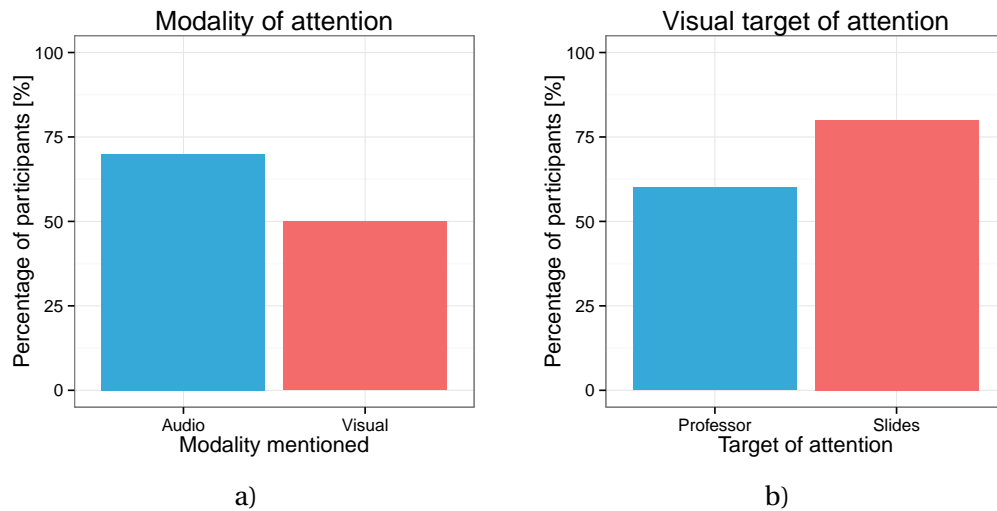


Figure 4.14 – Modality of focus emphasized by students a) based on the number of attributes used divided by their modality, b) visual target of attention mentioned.

immediate neighbour, 3 managed to recollect who was sitting in front and 2 remembered who was sitting behind them. No participant reported positive feelings about their neighbours, but 2 did have negative.

Other influences from the environment made little impression on the students (only 1 student described an element of the classroom not connected to fellow class-mates). This was surprising, given that during one lecture (after which 4 interviews were conducted) the construction site next to the classroom caused significant noise (classroom windows were open). After mentioning the episode in the interview, only some of the students vaguely remembered “something in the background”.

The last part of the interview was concerned with the impact the experiment had on the students. Overall, the perception of the experiment was neutral, with small number of people expressing positive or negative feelings about the experience (Figure 4.16a). Combining that with the information shown in Figure 4.16b where only a single person found the experiment overall disruptive, we conclude that the experiment did not have a big impact on students' behaviour. Most of the students (8 out of 10) acknowledged the experiment, usually remarking that they observed and understood that it was conducted. The in-class format was confirmed to be disruptive because of the interruptions of the lecture by the majority (6), but no negative feelings were shown.

After comparing the mean attention to the reports of experiment observations we found no significant correlation, but samples either in favour or against the experiment were very small.

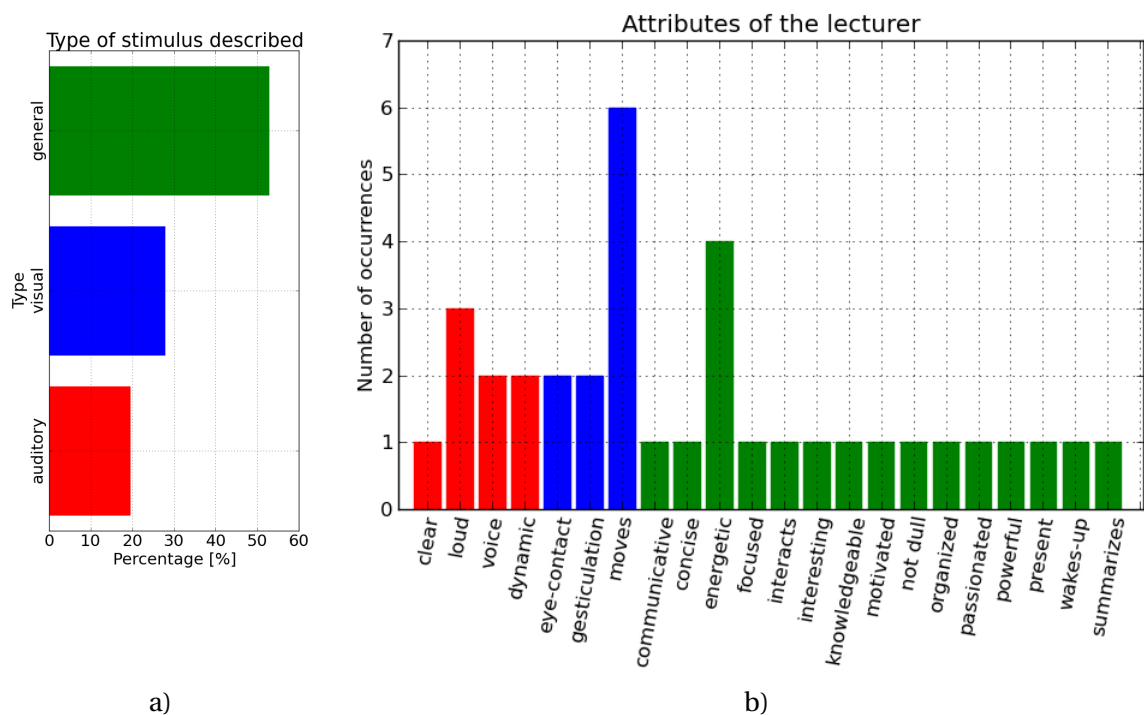


Figure 4.15 – Perception of lecturer based on the attributes used to describe his teaching a) divided by the type of attribute (percentage represents the ratio of all attributes used), b) list of attributes and their usage frequency (colours denote the type of attribute).

4.7 Conclusion

In this section we described our overall method for data collection of class perception from the students, focusing on the attention of the students. We compared the in-class and post-class formats and did not find significant differences on the perception of attention and perceived classroom attention, which were the two measures on which our research relies the most. We identified that attention of students does not fluctuate significantly during one lecture, but that it is changing from lecture to lecture.

We discovered that personal attention was an important influence on all other perceptions of the classroom. Similarly, reported activities of students correspond to expectations that lecture-related activities are dominant at the higher levels of attention. Our sample also showed that distracting activities were reported on almost all levels of attention.

We confirmed that attention is related with the distance from the teacher and that back rows have lower attention than the front rows. We also showed that the free-seating arrangement and student's mobility was not the major influence on this, and that students typically maintained the same location over all of our recordings.

Our interviews showed that the students overall did not have strong feelings about the experiment and that they did not feel threatened by the capturing apparatus, which leads us to



Figure 4.16 – Perception of the experiment by the interviewed students. a) Distribution of opinion about the overall experiment participation. b) Intrusiveness of the experiments in terms of how many subjects agreed with a certain statement.

believe that the recorded behaviour was natural. Although listening to the teacher remains the dominant form of transmitting information, we confirmed that students are aware of teacher's presence and gestural behaviour, which we see as a potential for our non-verbal measures (how much was the visual contact between the teacher and student maintained?).

5 Movement in class

Parts of this chapter have been published in Raca and Dillenbourg (2013) and Raca et al. (2014).

As we have seen from the classification in Section 2.3, human behaviour is rich with non-verbal expressions. In order to tap into the back-channel of human communication, our selection of features was guided by two attributes:

- unintentional - intentional signals, although needed and useful in a direct conversation, can be prone to conscious manipulation. Unintentional signals have the potential of being more truthful (discussed in Section 2.2.1).
- low entropy - as we have seen, individual gestures are known to be very informative, but narrowing down our focus too much can lead to low coverage of classroom situations. Signals of high entropy (clear meaning) such as head nods are an informative source of data, but easily interpretable without technological assistance and occur irregularly in the audience.

Intersecting two criteria, basic information which we can capture from a visual observation of audience is how much and/or how often does a person move, which will be the two main metrics discussed in this chapter. A low-level measurement such as this will be influenced by a number of physical and emotional factors. In order to differentiate signal from noise, our hypothesis is that:

- lecture provides a common signal which should influence all students to some degree;
- a large audience size will provide basis for an indirect synchronization (Section 3.1.4), which we can observe in motion of the individual.

By positioning the recording system in front of the classroom, we tried to minimize our intrusion on students' space. The setup and scenario did not come without challenges:

1. Camera's point-of-view recorded many occlusions and overlaps between students, caused by the dense seating arrangement. Our approach depends on visual contact with each individual.
2. Observed motion needs to be assigned to the person causing it, which was made difficult by the inter-personal occlusions. Each persons in the vicinity of motion registered by the camera needs to be evaluated as the potential source, and each motion needs to be connected to a single person.
3. Camera's location in front of the classroom made people in the back rows appear smaller due to perspective distortion. Developed measures should model person's behaviour independent from the distance, to the extent of technological limitations.

First problem, although manageable with additional cameras or dedicated recording room, was not solvable on algorithmic level. For our experiment we excluded students who were heavily occluded by the person sitting in front of them (e.g. overlapping bodies with minor surfaces of the rear person visible).

In order to handle the problem of motion assignment we introduced additional annotated data about student's location. Each person in the video is annotated with a rectangular region, and labelled with student's ID. Illustration of regions is given in Figure 5.1a, with the overlaps occurring between regions shown in Figure 5.1b.

We will devote part of this chapter in order to explain applied sanitation steps for solving problems 2 and 3, with majority of work dedicated to the motion assignment problem (which person is responsible for detected motion). Conclusions will be based on the smaller part of our dataset, and will include Class #1 (18 students) and #2 (38 students) each collected in a single recording session. For details about the sample, please refer to Table 4.1 and Section 4.2. Our observations will rely on Class #2 for validity, because of the bigger size of the population. We will also illustrate our findings on Class #1, and where possible we will combine the two classes to demonstrate the overall trend.

Our assumptions for the experiment are that students remain seated in their seats for the duration of our recording and that the camera is located in a fixed position.

5.1 Method

In its core, detecting motion in videos relies on finding the displacement of a distinctive point between two frames. We will briefly introduce the concept of *optical flow* as an established method used in our research. After that, we will give a detailed overview of additional processing steps needed for our scenario.

Our motion-extraction algorithm gives us a measure of motion intensity over time for each observed individual. The different steps of the algorithm are focused at ensuring reliable mo-

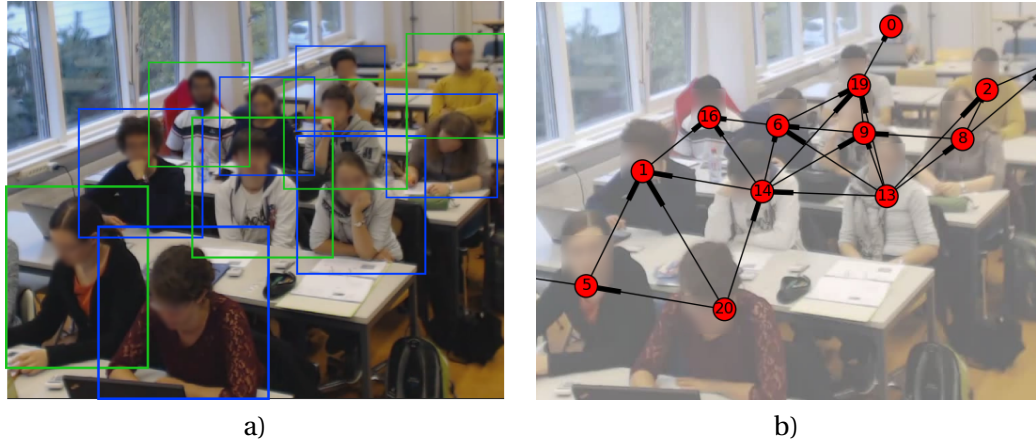


Figure 5.1 – Annotated student regions and overlaps. **a)** Rectangles represent the annotated regions connected with each student ID (different colours of rectangle's borders used for better visualization of occurring overlaps). **b)** Each edge of the graph represents an overlap between two regions and potential ambiguity for motion assignment. Direction of the edges is oriented towards the overlapped (occluded) student.

tion detection (outlier elimination, tracking-point distribution), motion assignment (“motion tracks”) and normalization which make the output comparable between students. Overview of the algorithm is given in Figure 5.2.

5.1.1 Optical flow

Forsyth and Ponce (2003) define **optical flow** as the “*motion of individual pixels in a video [which] is measured by attempting to find pixels in the next frame that correspond to pixels in the current frame (correspondence being measured by similarity in colour, intensity and texture)*”. Optical flow can be defined as a vector field of **motion vectors** estimated at **tracking points** of the image. Illustration of a simple optical flow is shown in the bottom row of Figure 5.3.

A dense optical flow will give us the “image velocity” for each pixel present in the image. This implies that observed transformations in 3D space can be represented as a vector field in the image plane. This is true for most situations and wide-spread implementations of optical flow can handle affine transformations (i.e. translation, rotation, scaling, homothety, etc).

For our studies we used the pyramidal implementation of optical flow (Bouguet, 2001) based on Lucas-Kanade tracking (Lucas et al., 1981; Lucas, 1985). Lucas-Kanade (abbreviated **L-K**) is a feature tracking principle based on spatio-temporal derivatives of the input image. The method was shown as one of the two most reliable in a comparative study done by Barron et al. (1994). Tested on a number of scenarios, the method’s mean angular error was below 1° on synthetically generated images and 4° in the worst case.

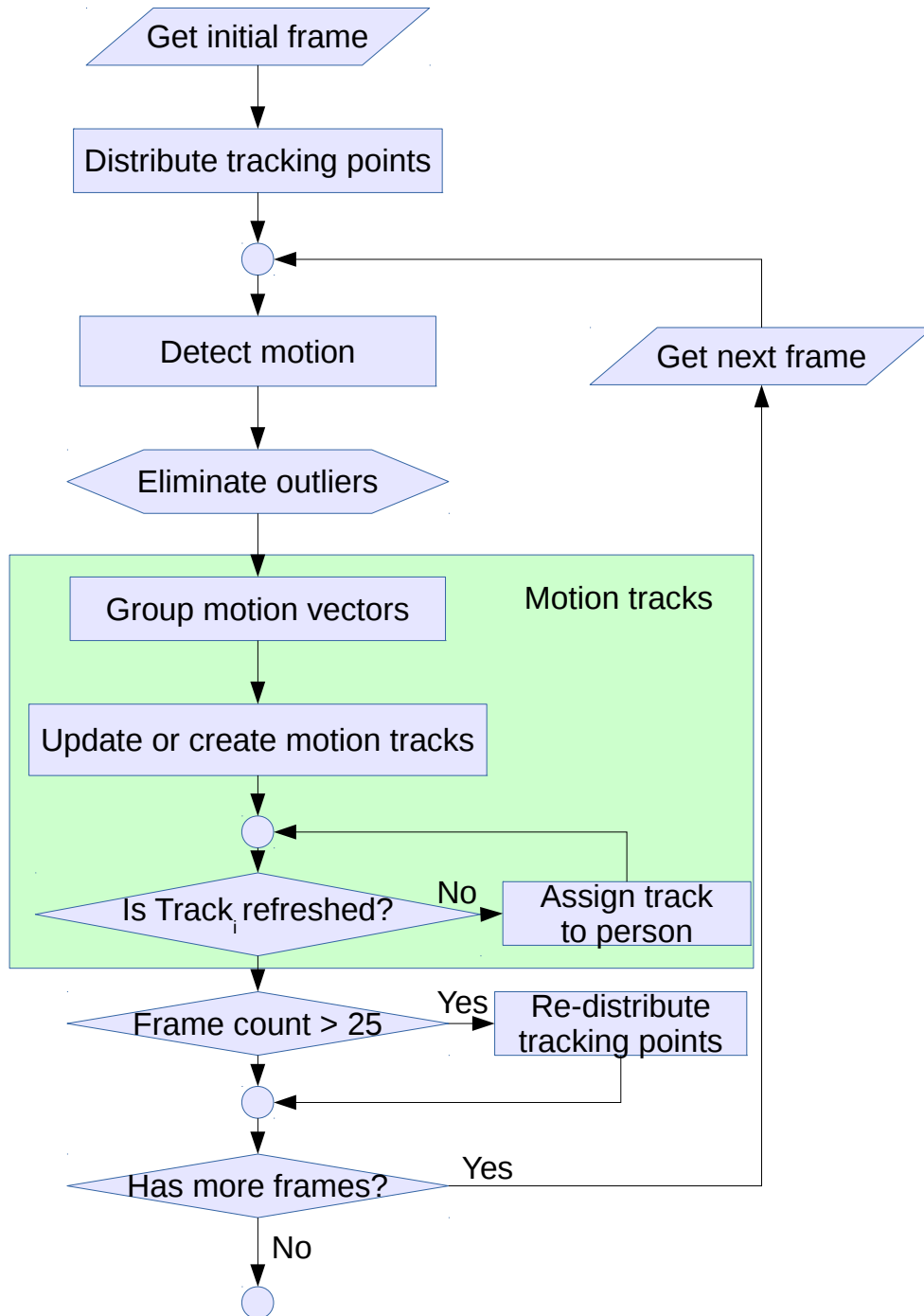


Figure 5.2 – Algorithm used for motion extraction and assignment. Overview of motion detection is given in Section 5.1.1. Distribution of tracking points is described in Section 5.1.2. Outlier elimination is detailed in Section 5.1.3. Motion tracks are described in Section 5.1.4.

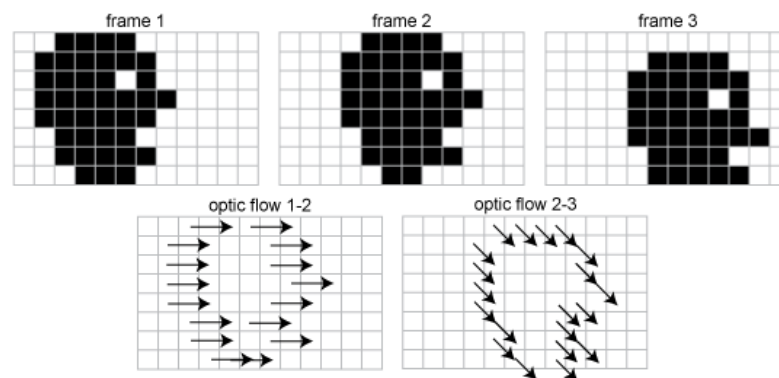


Figure 5.3 – Illustration of the optical flow. Individual image frames 1-3 are shown in the top row, while the difference between two respective frames is shown in the bottom. Optical flow is represented by the arrows representing motion vectors, shown for each pixel in the bottom row. Image copyright by Florian Raudies, Licence: CC Attribution 3.0.

Besides accuracy, the second desirable property of optical flow algorithm is robustness. Given that the local windows (regions of the image) for comparing points are relatively small (e.g. 20x20 pixels), algorithm can achieve most precise results if the motion of the points tracked is smaller than the considered window size. In order to make the algorithm robust to large movements, additional steps need to be applied.

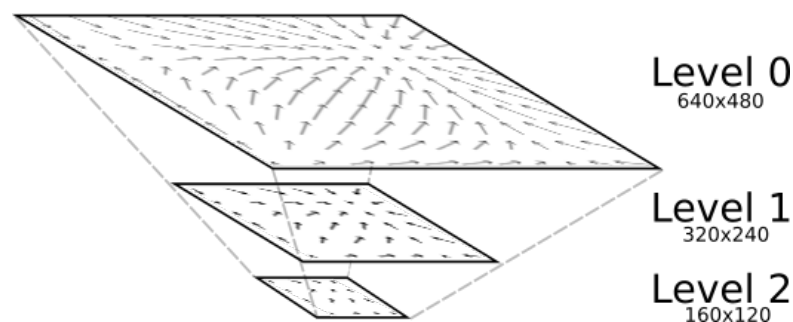


Figure 5.4 – Visualization of the pyramid down-sampling, 0th level representing the original image size, and each next level scaling down image size by half. Image sizes under the layer label are given for illustration purposes.

The pyramidal implementation proposed by Bouguet (2001) is based on creating a pyramid of down-sampled images (Figure 5.4). A frame is taken at its original size is the 0th level of the pyramid, and each new level is created by halving the image's dimensions. The procedure continues until a pre-determined limit is reached (typically, the pyramid has up to 4 levels). Optical flow analysis starts by estimating the motion vectors at the lowest level (smallest image size), where the large motion has been down-sized to a measurable value. Calculated vector field from the lower level is then used as the initial estimate for the next level, where motion vector values are refined.

Usually, estimating motion vectors for each pixel of the image is redundant and computation-

ally demanding. Also, problems can occur for regions which are uniform in colour because of their ambiguity. Instead of building a dense optical flow, we chose to use the L-K as a feature tracker. This means that we are pre-selecting salient points for tracking, and creating a sparser, but potentially more accurate optical flow. Points were selected by using the Shi and Tomasi (1994) criteria, in the OpenCV implementation (Bradski, 2000).

All of the previously presented research is concerned on raising the precision of tracking between two images (frames of a video). These fundamental improvements are significant in all usage contexts. In our specific case, we have seen the opportunity to improve the algorithm on higher levels by introducing additional consistency constraints, which we will detail in the following sections.

5.1.2 Distribution of tracking features

In order to make students from the front and back of classroom comparable, we needed to ensure that the motion in both cases is sampled with approximately the same precision. This problem was addressed at a number of levels, first of which being the number of points at which the motion was sampled.

We base our motion tracking on features (points and corners) which have a distinct appearance and provide good tracking results because of that. Because the algorithm has no depth information in the image, it favours people in front because of the bigger visual presence. In order to even the number of tracking points dedicated to each student, we introduced an additional constraint about the number of points per annotated region.

Each student region is expected to be populated with a grid of 8x8 equally spaced tracking points. The size of the annotated region and the number of tracking points expected gives us the radius between neighbouring tracking points (r_n) for that region. Distribution of tracking points is conducted in three steps for each region:



Figure 5.5 – Distribution of tracking points for the optical flow. Each red dot represent a point for measuring a motion vector. Notice the difference of densities for the students in front and in the back. In regions with no good tracking features detected, tracking points are arranged in a equally-spaced grid (visible in the lower right corner).

1. Shi-Tomasi points are found and used as the best choices for motion tracking.

2. Uniformly distributed (8x8 grid) tracking points are added.
3. Points from step 2 are pruned so that each tracking point has no neighbours in the r_n radius, maintaining the density of tracking points introduced by the constraint, and using the better tracking points from step 1 where possible.

Regions are processed from bigger to smaller ones, which is equivalent to considering people from front to back of the classroom. That way the overlap between two regions is pruned when the smaller region is being populated, using the smaller r_n . Visualization of the tracking point “distribution” can be seen in Figure 5.5.

In order to adjust to the changes in video, tracking points are re-distributed periodically every 25 frames (approximately every second).

5.1.3 Outlier elimination

Second constraint is aimed at eliminating errors in measurements. We observed that major errors with L-K tracking would either be *i)* isolated and/or *ii)* several magnitudes bigger than the surrounding measurements. Given that our observations are based on the intensity of motion, outliers such as this had the potential to lead us to wrong conclusion.

Filtering out outliers was based on two assumptions:

- given the adjusted density of tracking points, every meaningful motion will be detected by more than one tracking point;
- neighbouring tracking points should have motion intensity of approximately the same magnitude.

Neighbourhood region was declared as a circle with radius of 40 pixels centred on the tracking point. Each measurement would be eliminated if no other measurements were detected within the neighbourhood region, or if the measurement intensity difference was bigger than 2 standard deviations from the region’s mean.

5.1.4 Motion tracks and assignment

Finally, we improve the correctness of assignment by introducing spatio-temporal consistency with motion tracks. Individual motion vectors can be assigned to a person with a probabilistic approach. This does not take into account situations in which motion of one person temporally crosses into the region of another.

We assume that individual motion vectors will always be a part of a bigger, consistent motion. This is supported by the properties of human motion (direction does not change abruptly) and

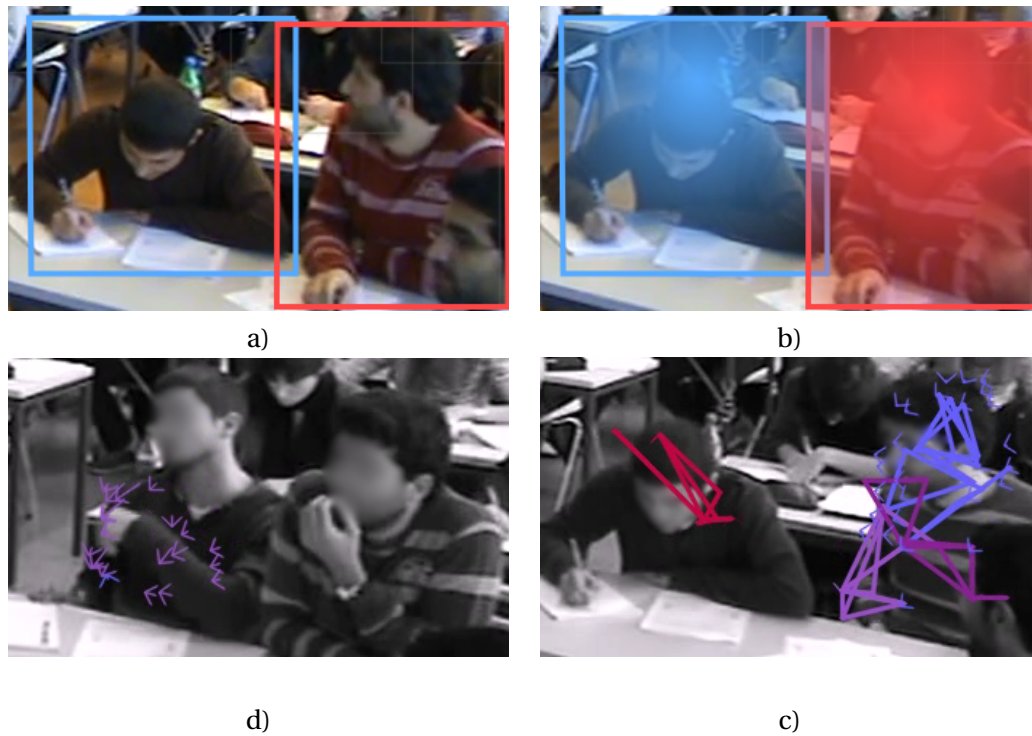


Figure 5.6 – Motion detection and grouping. **a)** Each annotated region is associated with a student ID. **b)** Marked student areas and illustrated centres of 2D Gaussian distributions which model the probability of motion belonging to a student. **c)** Individual motion vectors shown as purple arrows. **d)** Motion vectors grouped into motion tracks (represented as jagged lines) which can be assigned to an individual.

high sampling rate (motion typically lasts more than one frame). With this, we can increase assignment correctness by estimating the source of the entire motion, modelled by a motion track. **Motion tracks** represent connected “clouds” (spatially connected groups) of motion vectors over several frames.

In a single frame, motion vectors (\vec{v}) are spatially grouped with a variation of *region growing* algorithm:

1. Select a random unassigned motion vector, and initialize group's mean direction based on this seed vector.
2. Find neighbouring vectors within the predefined radius, and assign them to the group if the difference between their direction and group's mean direction is less than 45° .
3. If we were able to find vectors that qualify, update group's mean direction, expand the neighbourhood search radius and return to previous step until no new additions were found.
4. Repeat the algorithm for the next unassigned motion vector, until all detected motion

vectors in the current frame are assigned to groups.

After the groups in the new frame were created, they are connected into motion tracks (T) which spans across frames and models temporal consistency. Groups either initiate a new motion track, or are assigned to an existing track detected in the previous frame. Assignment is done with a *greedy* algorithm:

1. For each vector group in the current frame, calculate the group's centre of mass (mean value of (x, y) locations of tracking points in the group).
2. Merge the group with the motion track which has the closest centre of gravity in the previous frame, if the distance between the two centre is smaller than the specified threshold.

Raw motion vectors are shown in Figure 5.6c as purple arrows. For visualisation purposes, a set of cloud centres from several frames are connected into a track which is represented in Figure 5.6d as a coloured jagged line.

At each frame a motion track is either refreshed (a group of points was assigned to it), or is terminated. A terminated motion track is taken out of the pool of active tracks, and is evaluated to be assigned to a specific student.

The track is assigned to the student with highest probability of generating the entire motion (g_b), determined by Formula 5.1. Each student (g) is modelled with a 2D-Gaussian distribution centred on the the head location (depicted in Figure 5.6b). Function $p(\vec{v}|g)$ represents the probability of motion vector \vec{v} being generated by student (g), based on the location of the vector in respect to the 2D Gaussian associated with the student.

$$g_b = \arg\max_g \sum_{\forall \vec{v} \in T} p(\vec{v}|g) \quad (5.1)$$

5.2 Motion Analysis

Amount of recorded motion varied significantly from person to person. In captured units, each motion vector represents the distance of motion displacement in pixels. Total motion intensity for a given person was captured by summing intensities of all motion vectors within the time step (2 seconds, explained in Section 5.2.3). In order to create a measurement which we can compare between individuals, normalization step was required.

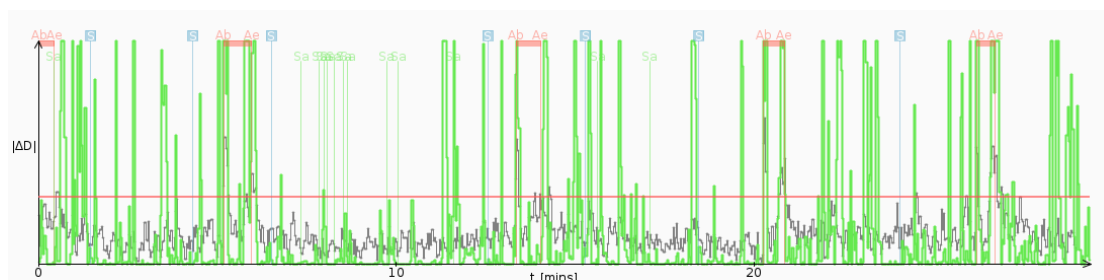


Figure 5.7 – Visualization of motion intensity for a single person (green) over the class mean (grey). Horizontal axes represents the time, vertical the relative intensity of motion for the given person. Vertical markers represent the annotated classroom events, most notably - the 4 paired red markings represent the moments of start and end of the questionnaire fill-out. Horizontal red line shows the minimal intensity of motion considered an “observable movement”.

5.2.1 Motion normalization

Our distribution of motion-tracking points ensures that we provide the same chance for detecting motion of people in the front and back, but does not guarantee the same motion intensity for equivalent actions. In order to achieve this, all vector intensities associated with a specific student are normalized by the diagonal length of the student’s annotated region.

Given that we do not consider the nature of the motion, our final aim is to transform the observed motion intensity to **relative motion intensity** which we can relate to the person’s in-class activity. The final range of relative motion intensity is between 0-100% and is based on two observations from our recordings:

1. Student is on average sitting still during the class.
2. Student has at least one full upper-body movement in the recorded footage (e.g. pose shift).

In accordance with these assumptions, our scaling mechanism also consists out of two steps:

1. We take the median value of movement intensity as the 5%. The value was taken heuristically to approximate small motion with a reasonable value, and allows us a large scaling space for bigger movements.
2. The algorithm checks that given the 5% motion intensity value, the student reaches 100% motion at least once during the class. Motion which registers above the threshold of 100% is clipped to the maximum value.

The *relative motion intensity* measure for a single person is shown in Figure 5.7. We acknowledge that this scale is tightly connected to our scenario (students sitting and taking notes, with occasional body-shift) and that the scaling should be re-defined in contexts which include a more dynamic behaviour.

5.2.2 Observable and synchronized movement

Given that 100% of relative motion intensity is roughly equivalent to full upper-body movement, we defined **observable movement** as motion with more than 30% intensity. The 30% threshold was heuristically taken as the limit which separates minor body movement and motion that can be registered by people in the student's surroundings. The motion intensity is visualized in Figure 5.7 as the horizontal red line.

Synchronized movement instance between two persons is defined as two instances of observable movement happening at approximately the same time. We use the term “approximately” because we consider different time delays between the motions to model different synchronization precisions. Detailed explanation is given in the following section.

5.2.3 Motion synchronization

From the dual eye-tracking theory, we know that quality of collaboration (Richardson and Dale, 2005) and understanding (Jermann and Nüssli, 2012) between two persons can be assessed by analysing the correlation of their gaze patterns. In our work, we expand on previous conclusions in two ways - *i*) by analysing the whole audience and *ii*) using a more general measure of activity. Our hypothesis is that students who listen to the teacher will be more likely to move in a synchronized manner, while an absent-minded student will act on his/hers own internal rhythm.

Synchronized motion is not limited to a specific action, but can be explained on example of note-taking - attentive students would turn pages on the hand-outs and take notes immediately after they were presented in class. More than a reaction to lecture's audio/visual stimulus, motion can be seen as a “convergence” of audience, or *indirect synchronization* to a signal (Section 3.1.4). If students perceive an outside event (e.g. loud noise, truck) as more important than the lecture, they would still have a synchronized motion (everybody looking through the window) but caused by a different stimulus than the teacher. In their publication Delaherche et al. (2012) terms this concept as *process coordination*.

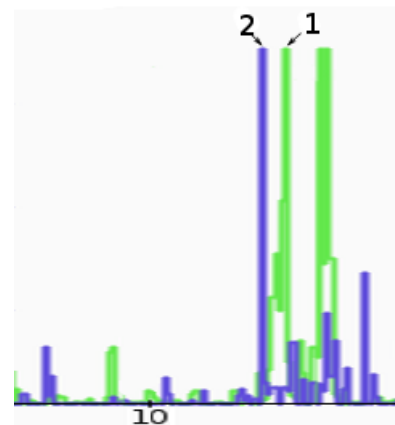


Figure 5.8 – Example of co-movement on a movement intensity graphic of two persons. The picture is a snippet of a motion-visualization as shown on Figure 5.7. We are displaying motion intensity of two persons overlayed over each other. Person 2 shifted hers seating position (blue line), 2 seconds later, neighbouring Person 1 (marked in green) also started re-adjusting herself.

Classroom synchronization was studied in a dyadic fashion, by comparing pairs of any two students. We used only the data collected during uninterrupted teaching, and did not take into account the questionnaire fill-out periods.

Similarly to the analysis of synchronization between pairs done by Delaherche and Chetouani (2010), we took into consideration the seating arrangement between two analysed students. We divided the dyads into three conditions based on their mutual visibility: immediate neighbours, visible neighbourhood or non-visible student pairs (as described in Section 3.2). In the case of immediate neighbours, the students were considered as mutually visible (both of them can observe and synchronize to the actions of the other). In case of visible neighbourhood, student sitting behind can observe the actions of the other student, but not the other way around. Non-visible student pairs were considered for cases of accidental and indirect synchronization, but not as a direct influence of one on the other.

Given that learning is not a scripted activity, reactions of students can vary or be completely blank. The research in dual eye-tracking has identified a delay of 2 seconds between the speaker's and listener's gaze when a specific item was referenced (Richardson et al., 2007). The conclusion of that research was that the comprehension between participants is inversely proportional to the time-lag. Based on this threshold, we define actions of two students as **co-movement** if the actions co-occur within a time window of 4 seconds (depicted in Figure 5.8). We differentiate between:

- perfect synchronization, < 2 seconds apart,
- synchronization, 2-4 seconds apart (2 seconds added as time needed for executing the motion),
- weak synchronization, 4-6 sec apart.

Third period (4-6 seconds) was introduced to take into account mimicking - when the person is not reacting to the teachers stimulus but is following the reaction of others, in which we add 2 seconds for the person to observe the reaction of others and then reproduce it. **Sleeper's lag** represents the delay ("lag") in movement caused by mimicking actions of other students instead of reacting to the original source of information.

Algorithmically, motion synchronization between two persons was calculated as matrix multiplication. Each person is represented with a time series of motion intensity values, sampled in 2 second steps. **Co-movement matrix** is created by multiplying the two time series as $N \times 1$ and $1 \times N$ matrix (visualized in Figure 5.9a). N represents the number of samples collected for each person during the lecture.

Within the two time series, values with the same index represent same time period in the lecture. This means that *perfect synchronization* moments will be found on the diagonal of the co-movement matrix, coordinates (t, t) . To analyse *synchronization* instances (2-4 seconds

apart), Person A who moved before will occur 1 time-step before, and the co-movement with Person B is located at coordinates $(t - 1, t)$. Similarly, “weak synchronization” with Person A moving 4 seconds before Person B is shown at coordinates $(t - 2, t)$. In the cases of mutual visibility, reverse direction of influence (Person B moving before Person A) is also possible and shown at coordinates $(t + 1, t)$ and $(t + 2, t)$.

Majority of the co-movement matrix represents synchronized movement instances which are too far apart from each other to be relevant (bigger difference between coordinates represents bigger time delays between actions). For that reason we focus on the diagonal and the two bands around it: $\pm 2sec$, $\pm 4sec$. From a perspective of Person B, we can densely represent synchronization moments with Person A as a time-line shown in Figure 5.9b.

Because the values in the co-movement matrix represent multiplication of motion intensities in the range (0.0 - 1.0), the value produced will be high only if both movements were of high intensity.

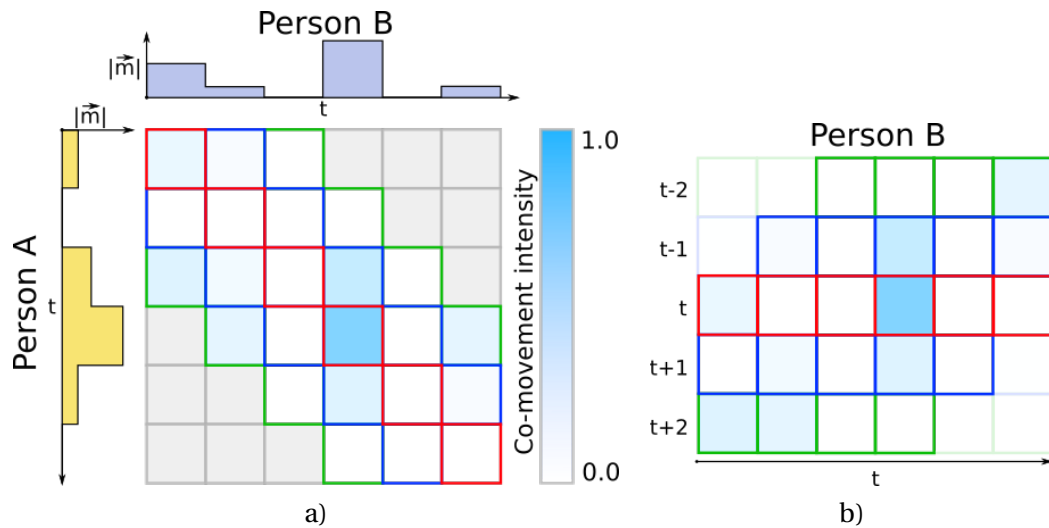


Figure 5.9 – Synchronized movement. **a)** Co-movement matrix of Person A and B over a period of 12 seconds (6 time steps). Perfect synchronization is represented by the diagonal of the matrix, marked with red squares. $< \pm 4$ second synchronization is represented with blue cells and weak synchronization ($< \pm 6$ seconds) is marked with green cells. Periods which were too far apart to be considered are grayed-out. **b)** Co-movement timeline, considered from the perspective of Person B. The figure shows the same values as the co-movement matrix, aligned on the diagonal cells of the matrix (red squares). Transparent sections are not present in the example matrix.

5.3 Research questions

We had a number of questions which we wanted to explore with the motion metric. The main questions can be broadly divided into analysing the effects of the spatial arrangement, and on the relationship between motion and attention.

1. **Spatial relationship between students** - Is there a detectable influence of the students on each other? We have sectioned the classroom space into a number of discrete categories depending on the relative arrangement between student. Can we see the mirroring of proxemic zones in the domain specific arrangement such as the classroom?
2. **Arrangement of students in the classroom** - In Section 4.5 we have shown that students typically choose the same seating location. Also, the questionnaire data has shown an negative influence of distance on attention. Is there an influence of the distance in the classroom observable in our motion measurement?
3. **Indirect synchronization** - Based on the general observation such as the amount of motion detected, can we formulate a measurement of indirect synchronization in the audience? If so, can this measurement be used as an indicator of student's attention?

5.4 Results

Question 5.3.1 We compared the average number of synchronized movements between pairs sitting immediately next to each other and other two pair types. We found that immediate neighbours had higher number of synchronized motion instances than a non-neighbouring pair with a t -test (significance $p \leq 0.05$). Results are shown in Table 5.1 and visualized in Figure 5.10.

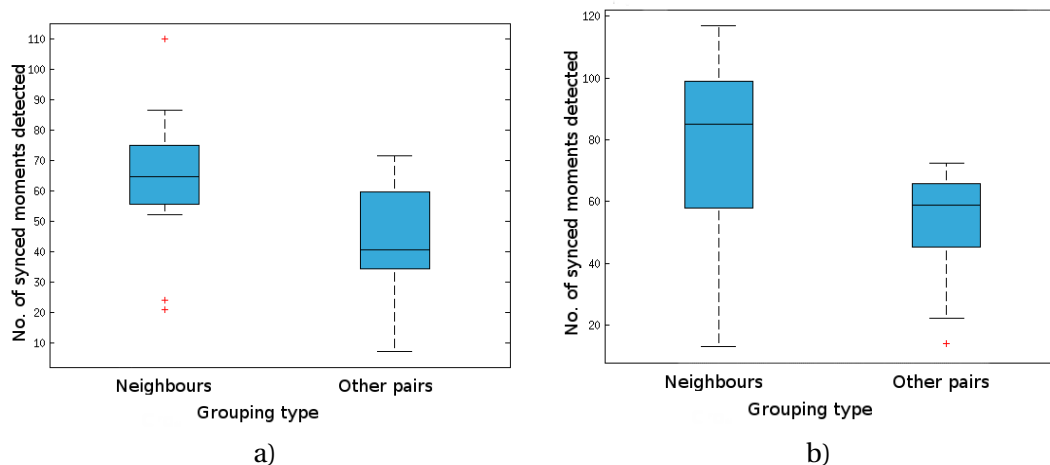


Figure 5.10 – Difference in the number of observed synchronized moments between pairs of neighbouring students and other pairs of students. Values shown for a) Class #1 and b) Class #2. Exact values shown in Table 5.1. Boxplots show the median value, the edges of the boxes are the 25th and 75th percentile and whiskers depict the 90% of the sample.

We found no significant difference in number of synchronized moments between a pair of students from the visible neighbourhood and a pair of students without visual contact. With this we can only report that we found no evidence that people in the visible neighbourhood influence the behaviour of a student in a detectable way.

Class	Neighbours $\mu (\sigma^2)$	Other pairs $\mu (\sigma^2)$	DoF	T-value	p-value
1	63.3 (24.3)	44.9 (18.4)	23	2.13	0.0424
2	76.5 (32.4)	54.4 (15.7)	53	3.55	0.0008

Table 5.1 – Average number of synchronized moments between immediate neighbours and other pairs, and results of the *t*-tests.

Question 5.3.2 To compare the motion metrics with the previous findings on student activity, we also tested the influence of teacher’s proximity to the movement of the students. “Total motion intensity” measure used here is calculated as a simple summation of all motion intensities captured for a single person. Distance is presented in unit measures, where each seat and row is represented as 1 unit of distance. Walking spaces between rows are also taken into account as 1 unit.

The further away students are from the centre-front of the classroom the less active they are, judging from the number of observable movement instances (Kendall correlation is $\tau(27) = -0.284$ ($p = 0.03$) for Class 2; and $\tau(16) = -0.172$ ($p = 0.45$) for Class 1). Analysing the samples we have seen the same trend in both cases, even though the correlation was insignificant for the first classroom. Figure 5.11 shows the correlation for Class 2.

This confirms observations of Adams (1969), which showed that front-centre of the classroom is responsible for the majority of interactions, and that they decline with the increase of distance from that point. The activity of individuals is not only present in the formal interactions, but is also reflected in the overall movement activity.

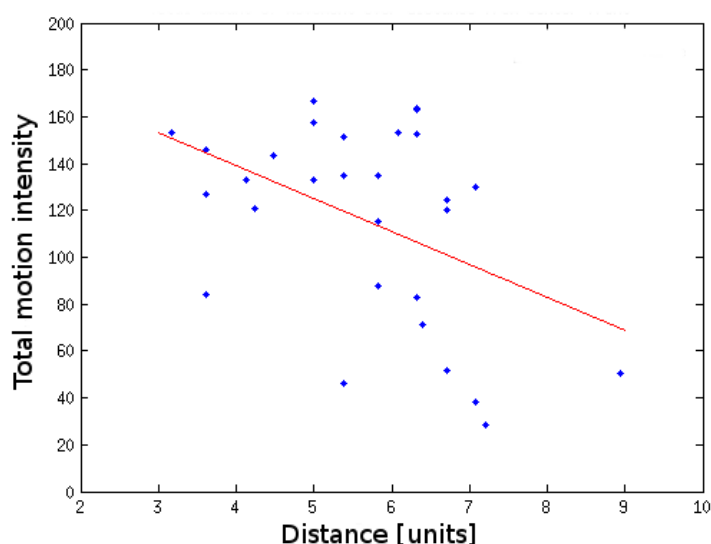


Figure 5.11 – Correlation between distance from teacher and motion intensity in Class 2; Kendall correlation $\tau = -0.284$ ($p = 0.03$)

Question 5.3.3 Our third test was to find the correlation of the average reported level of attention to the speed of reaction. The question was whether students with lower attention levels were more likely to lag behind a other students in their visible field.

Motion lag for person i (M_i) was calculated as a mean lag in synchronized moments of one person compared to the visible class population for that person $C(i)$. Visible class population includes the immediate neighbours and the students in the visible region.

Synchronization between two persons can be registered with a delay of $\delta \in \{1, 2, 4\}$, depending on the time between the their actions. Number of synchronized instances over the lecture period for a specific pair of person (i, j) and a given time difference δ is defined as $s_{i,j}(\delta)$.

The total motion lag is a weighted average of observed synchronised motion instances over the entire duration of the class

$$M_i = \frac{1}{N} \sum_{j \in C(i)} \left(\sum_{\delta \in \{1, 2, 4\}} \delta \cdot s_{i,j}(\delta) \right) \quad (5.2)$$

where N is the normalization factor, equal to the total number of observed motion instances for all parameter combinations.

The correlation found had the expected trend in Kendall correlation $\tau(27) = -0.259$ but was marginally insignificant $p = 0.06$ on the sample size of 29 students of the Class #2. The result is shown in Figure 5.12a. Class #1 correlation had a similar trend but was not statistically significant $\tau(16) = -0.222$ ($p = 0.32$). Combined data from both classes is shown in Figure 5.12b.

The data suggests that the phenomenon of “sleeper’s lag” exists, but the current sample is not conclusive. Also, the difference in average speed of reaction is in sub-second intervals, which makes the indicator observable only with the assistance of technical devices.

5.5 Conclusion

In this chapter we introduced general motion as a usable measure of classroom behaviour. We detailed the processing steps in order to present the potential technical pitfalls we observed, and how we tried to avoid them in the data sanitation stage. It is our conclusion that Computer Vision techniques have significantly improved in the general domain, but that additional improvements can be achieved by implementing context-specific steps.

We presented three conclusions in accordance with our theoretical goals. We demonstrated, by counting motion synchronization instances, that neighbouring pairs of students have an influence on each other. Considering the scenario of non-collaborative learning, this

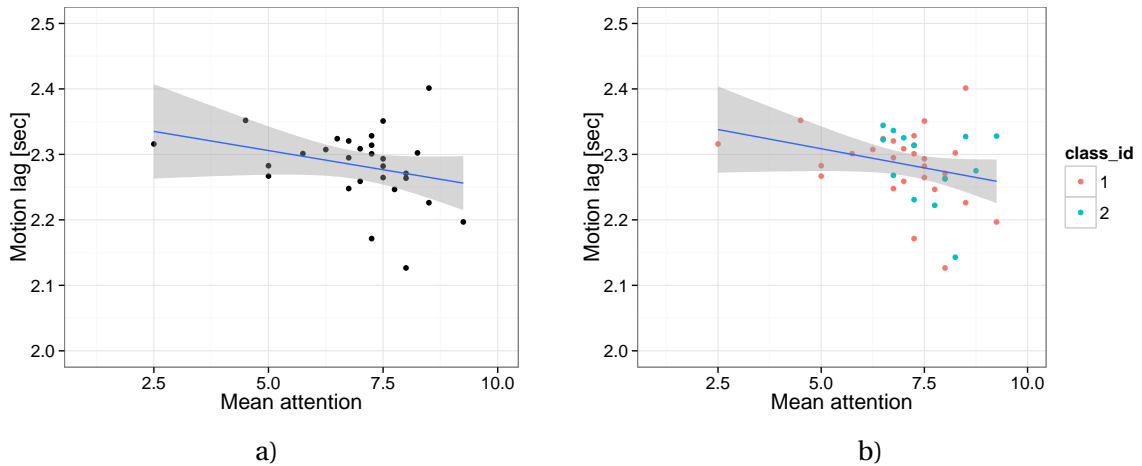


Figure 5.12 – Motion lag compared with the mean level of attention. a) Class 2 shows a Kendall correlation $\tau = -0.259$ ($p = 0.06$). b) Both classes combined show weaker correlation, Pearson's $r = -0.011$, ($p = 0.09$). Colours of data points represent Class 1 and 2 samples. Gray regions of fitted lines in both cases represent the 95% confidence interval.

potentially opens new discussions about the benefits of seating arrangements and neighbour selection.

Our second conclusion re-affirmed the influence of spatial arrangement on students, observed in other studies with a different set of measures. Similarity between conclusions reached with our methods and human observations is a welcomed validation of the novel approach.

Finally, we defined the measurement of “motion lag”, and showed that it has the potential for detecting under-performing individuals. The principle on which this indicator is based corresponds to classroom entropy – the bigger the “chaos” in the motion of students, values measured will be more randomized (based on the purely accidental co-movements). In the opposite direction, with a steady based of “tuned-in” individuals, the metric has the potential to step around behaviour mimicking and identify the true state of attention in the classroom.

Our conclusions suffer from a small experimental proof. The samples collected and processed point in the correct direction, but further tests should be performed to find the optimal capturing setup and precision of measurements.

6 Participation of head and eyes

GAZE-based measurements have gained popularity in the both scientific community and marketing research for their precision and highly informative output about subject's intentions. Unfortunately, the technical constraints (expensive, intrusive devices) do not allow for the wide-spread usage of these measurement. The natural approximation of gaze information that can be captured from a distance is the orientation of the head.

In this chapter we will show how practice of head-tracking has been used before for estimating gaze information in meeting settings, and we will expand on the existing research in order to create a model which we can further use in the classroom context. The focus of our investigation will be on the horizontal head/eyes rotation (yaw), as the dominant orientation in everyday usage. The main goal of the chapter is to increase reliability of our behaviour predictions by accurately modelling eye's behaviour from head orientation observation.

6.1 Previous work

The major overview of eye-tracking methods by Holmqvist et al. (2011), specifies that the visual range of human sight is $\pm 40^\circ$ horizontal and $\pm 25^\circ$ vertical (page 58). Spector (1990) gives a more detailed specification by putting the ranges of the field of view for each eye to 60° vertically up, 75° vertically down, 60° horizontally inwards (towards the nose) and 95° horizontally outwards. Other research demonstrated that human eyes have a limit of around 45° of freedom, imposed by neural mechanism, and not physical properties of the eyes (which limit the range to around 55°) (Guitton and Volle, 1987). Unfortunately, neither of the sources gives us information on the actual usage of the range.

Overview of eye-movement in natural behaviour by Hayhoe and Ballard (2005) noted that eyes are guided by an internal reward system, and that the eyes are not positioned on the most salient target but in the position best adjusted for the given task. This implies that different activities will show different gaze patterns and that our tests need to be similar to the field of application we have in mind.

In a simplified formulation, Yücel and Salah (2009) set the gaze uncertainty to constant value of approximately 8.43° , determined by finding a trade-off between the observed gaze target precision and number of objects in the experiment. The threshold is influenced by the proposed scenario, in which a group of visually salient objects is placed on the table in front of a user. It is important to notice that the value is only slightly larger than the implicit uncertainty of the gaze (Pöppel and Harvey Jr (1973) specified the *fovea centralis* as $\pm 2^\circ$, which we consider as the main channel for information reception, followed by *perifovea* region, $\pm 10^\circ$ wide). This implies that people will move their eyes approximately $\pm 6.43^\circ$, or in only 15% of the available range.

In meeting situations, Stiefelhagen and Zhu (2002) tested the idea on the scenario of 4 persons sitting around the meeting table, with the usage of a head-tracker. Results showed that head orientation contributes 68.9% in the overall gaze direction (where is the attention directed) and achieved 88.7% accuracy at determining the focus of attention between the 3 possible targets. The approach was later criticized for over-fitting on a small data-set (Ba and Odobez, 2006) and turning the problem into classification between 3 possible targets (Voit and Stiefelhagen, 2008).

Improvements were suggested by offering more discrete targets (people, projectors and table-top) and creating a probabilistic mapping between head-orientations and potential targets (Ba and Odobez, 2009). This work also proposed a geometrical model of head participation defined as

$$\alpha_H = \begin{cases} \kappa_\alpha \alpha_G & \text{if } |\alpha_G| > \epsilon_\alpha \\ 0 & \text{if } |\alpha_G| < \epsilon_\alpha \end{cases} \quad (6.1)$$

where the head angle (α_H) is in a linear relationship with the overall direction of the gaze (α_G) after a certain threshold is passed (ϵ_α), below which we assume that the movement is carried out only by eyes. The parameters used in the paper assumed that the relationship is linear for all head rotations ($\epsilon_\alpha = 0$) and that the α_H assumes values from the range $[0.5 - 0.8]$ which were individually fitted to each subject. The model depended on a per-user and per-location parametrization in order to improve the precision, which makes the conclusions difficult to scale to general population.

Voit and Stiefelhagen (2008) found the value of $\kappa_\alpha = 0.72$ (which represent the participation of head in the combined gaze direction), based on the annotations of subject's target of gaze made by human evaluators. The conclusion suffers from two problems – *i*) due to a discrete set of targets, the range of head motion is not evenly represented, and more importantly, *ii*) the conclusions suffered from the ambiguities of targets and potential annotation errors. This paper modelled the viewing frustum (field of view) with a cone of 60° horizontal and 50° of vertical width.

The linear model used in meeting scenarios was based on research carried out by Freedman and Sparks (1997). The tests were conducted on two trained rhesus monkeys, and found that



Figure 6.1 – Portable eye-trackers. Two of the popular brands are the **a)** Tobi Glasses and **b)** SMI portable eye-tracker, shown here with the SMI Optical Head Tracking module. Images copyrighted by the mentioned companies.

the participation of head was visible at angles above 20° , and followed a linear trend. At larger angles, eye participation saturated around 30° .

In a complex meeting scenario, Voit and Stiefelbogen (2010) tried to estimate the gaze in the environment which included objects, moving and static human targets, with the usage of volumetric representation for gaze targets in order to account for occlusions. The conclusion of the work was that in a complex situation, human gaze can not reliably be modelled just by head orientation, and that other contextual knowledge is needed for determining the target of attention.

6.2 Methodology

With the new generation of portable eye-trackers (shown in Figure 6.1), we were able to both carry out several captures of human gaze patterns in real-world scenarios, and in controlled experiment settings. In our experiments we used the SMI portable eye-tracker, without the head-localization add-ons (shown in Figure 6.1b).

Because the output of the SMI software package was the location of the gaze/pupils in the recorded image of the eye-tracker, we needed to convert the output to angles in order to process the data, and determine the limits that the eye-tracker imposes on the user.

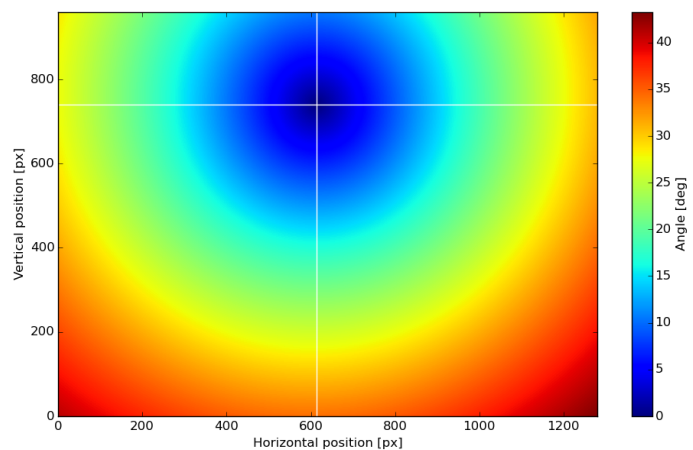


Figure 6.2 – Undistorted view angles of the eye-tracker in degrees, shown for every location of the recorded view-field. Centre is marked with the white horizontal and vertical lines to display the horizontal and vertical offset.

We measured the *field of view* of the eye-tracker by positioning it on a fixed post in front of a whiteboard with a drawn grid. The recorded field of view spans 27.7° to the left, 30.2° to the right, 11° upwards and 34.9° downwards. Although the centre of the view-field rested in the horizontal centre of the image, there was a notable shift in the vertical direction (shown in Figure 6.2). This did not have a major effect on the tests, because the primary focus of our research is in horizontal angles of eye-motion.

Intrinsic parameters of the frontal camera (used to un-distort the recorded image and coordinates of the gaze) were acquired by showing the check-board pattern in front of the eye-tracker and using the standard OpenCV library methods (Bradski, 2000) for both finding parameters and un-distorting the images/coordinates.

6.2.1 Real-world usage

For the real-world scenario we chose to give the eye-tracker to teachers conducting the lectures in our experiments. The subject is thus free to move with the implicit assignment of scanning a large group of people with wide spatial arrangement (format of the classes were lectures, not discussions). This meant that the eye-movements will be active in a wide range of unstructured movements for time periods of 35-40 minutes. An example of teacher's point of view is shown in Figure 6.3a. In total, 6 recordings were captured.

The subjects were introduced to the eye-tracker before the experiment. The device was adjusted to the user with a 3-point calibration at the start of the lecture. No specific instructions about the behaviour were given, apart that the subject should behave naturally.

6.2.2 Controlled experiment

Controlled experiment aimed at capturing the head/eye configuration of people looking at a single stimulus at different horizontal angles, while the stimulus was kept in the same height with participants' eyes.

In order to reach a reliable sample we had 3 independent variables in the experiment:

- the **horizontal angle** of the stimulus. By positioning the subjects near the projection plane (approximately 70cm), we were able to simulate horizontal angles up to $\pm 45^\circ$. The angle range was sampled in discrete steps of 5° , so the available angles used were 0° , $\pm 5^\circ$, $\pm 10^\circ$, $\pm 15^\circ$, $\pm 20^\circ$, $\pm 25^\circ$, $\pm 30^\circ$, $\pm 35^\circ$, $\pm 40^\circ$, $\pm 45^\circ$;
- the **travel** of the stimulus - to simulate if head/eyes pose would be different depending on how the pose was reached, we simulated either a slow-moving stimulus (the stimulus moved into the angular position by slowly sliding, known as “**smooth pursuit**”), or a sudden-relocating stimulus (the stimulus “**jumped**” to the angular location with no transitional animation).

Angle	5°	10°	15°	20°	25°	30°	35°	40°	45°
Travel	J	S	J	S	J	S	S	S	S
Stay (sec)	10	10	5	10	20	5	10	20	5

Angle	-5°	-10°	-15°	-20°	-25°	-30°	-35°	-40°	-45°
Travel	J	S	J	S	J	J	S	S	S
Stay(sec)	20	10	5	20	10	5	20	10	5

Table 6.1 – Used combinations of horizontal angle, type of travel (*S* - smooth pursuit, *J* - jump) and duration of stay in seconds.

Stay duration	5 sec	10 sec	20 sec
Jump	-30°, -15°, 15°	-25°, 5°	-5°, 25°
Smooth pursuit	-45°, 30°, 45°	-40°, -10°, 10°, 20°, 35°	-35°, -20°, 40°

Table 6.2 – Division of angles used between the stay duration and transition type.

- **duration of stay** - to capture how/if the pose changed over time, after reaching the defined angular location, the stimulus would remain there for either 5, 10 or 20 seconds.

All combinations of 3 independent parameters required a considerable amount of time and effort from the subjects and were not manageable for this study. Our primary interest in the experiment is the influence of the horizontal angle on the head/eye pose, with the time varied for preventing the subjects to anticipate the next event in the experiment. Two travel styles were also introduced in order to simulate different situations of gaze usage in everyday life.

The compromise was to vary the *travel* and *duration of stay* parameters for each angular location in order to sample the combinations as much as possible. The used combination of parameters in the experiment is shown in Table 6.1.

The values of *stay* durations were cycled for each angular location (cycling through the three values for each next angular position). The *travel* condition was originally following the same principle, but after the pilot studies we realized that the participants were losing sight of the stimulus if it “jumped” to a position over 30° due to the eye-tracker’s field of view. To avoid confusing the subjects, all angles above 30 degrees were put in the *smooth pursuit* condition (leading to the dis-balance between the two conditions).

Table 6.2 shows the division of angles among the 6 possible combinations of travel and stay duration. Even though the dis-balance exists, this arrangement allowed us to have good coverage of the angles for each duration of stay and type of stimulus travel. The duration of the experiment was little over 5 minutes, which we considered optimal for a high-concentration task (with calibration and instruction steps, the entire experiment lasted around 25 minutes).

To neutralize the influence of the order in which the angles were shown, after each position the stimulus was returned at the centre of the subjects field of view (0°) for the duration of 3

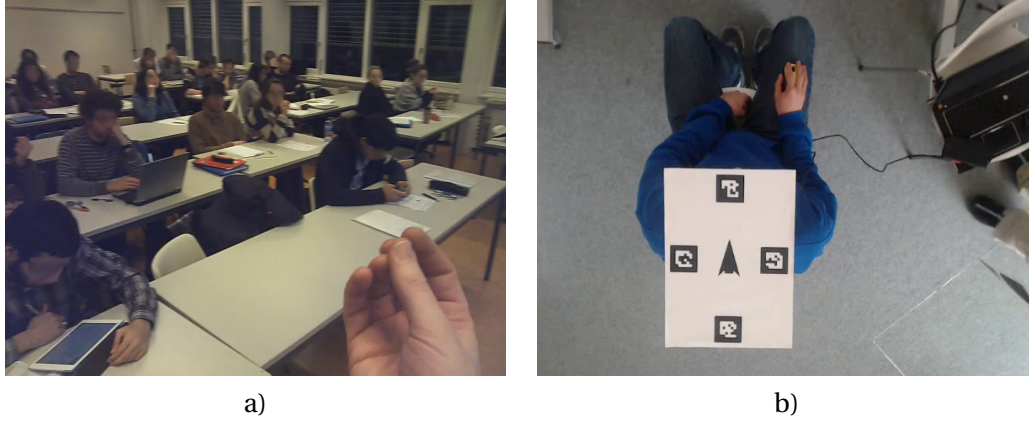


Figure 6.3 – **a)** Teacher's view from one of the recorded classes. **b)** Top-view of the controlled experiment showing the seated student wearing the head-pose tracking markers (chili-tags) on his head. Recording and synchronization devices visible on the right half of the image.

seconds. The order in which the angles were shown was also randomized to prevent subject from anticipating and “preparing” for the next position.

6.2.2.1 Capturing methodology

For the gaze recording we used the same undistorted and calibrated output from the SMI eye-tracker mentioned in the previous section. Head orientation was captured by placing a surface with a set of markers (Chili-tags Bonnard et al. (2013)), shown on Figure 6.3b. We also explored the option of using Kinect device to capture the head orientation unobtrusively, but the measurements acquired from the Microsoft SDK had unacceptable error sizes for the experiment.

We captured 3 streams of data (visualized in Figure 6.4):

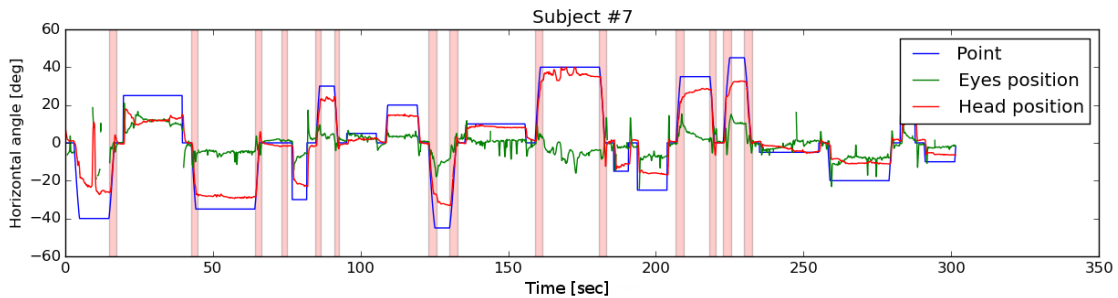


Figure 6.4 – Visualization of the data captured in the controlled experiment. The three streams of data shown are the ground truth of the point angular position (α_{GT} , blue line), horizontal angle of the eyes (α_E , green) and head (α_H , red). The vertical red blocks were the periods when the stimulus was not in the “observed” state. Shorter unobserved periods within a single angular fixation are not visible due to their short duration.

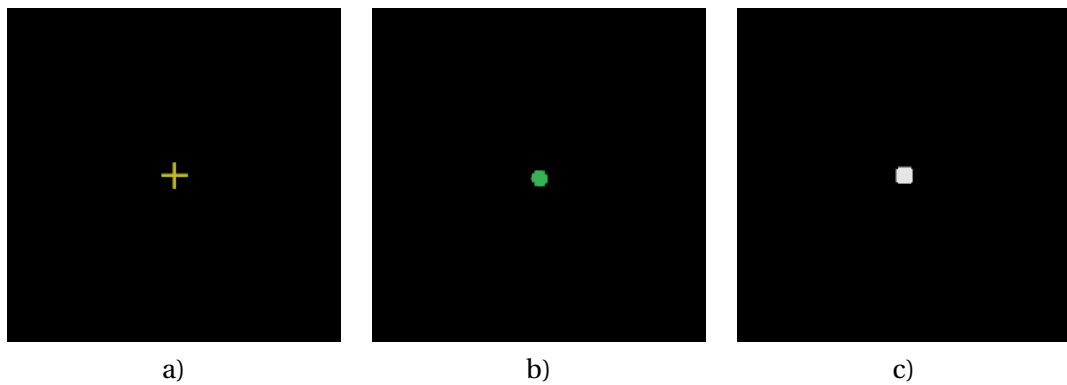


Figure 6.5 – Visual stimuli shown to the subjects during the controlled experiment. **a)** The stimulus would initially be shown as a yellow cross for easy localization by the subject, **b)** after the user has clicked the mouse button (indicating that he sees the item), the item would change its shape and colour to green circle to indicate that **c)** when the item was moving, it would be displayed as the white square to enable easy tracking by the subject.

- the eye-tracker (video of participant's view-field and the pixel coordinates of the gaze position within it). From the eye-positions in the view-field we extracted the horizontal angle of the eyes (α_E);
- the head-pose tracker (from the camera located above the subject, shown in Figure 6.3) - capturing the locations of the 4 markers placed on the head of the subject, from where we extracted the horizontal angle of the head (α_H) by simple geometrical calculations;
- the position of the stimulus on the projection surface, more precisely - the horizontal angle (α_{GT}). The application responsible for projecting the stimulus image was calibrated based on the subjects height and distance from the projection surface, which allowed us to record the spatial location of the item being tracked, and the angle the item had in respect to subject's head. We use this stream as the ground-truth.

All streams were synchronized with millisecond precision in the post-processing steps. Data was recorded in different time resolutions: eyetracker and head-pose tracker recording at 30fps, while the ground-truth was captured at 10fps (the speed was limited by the system performance, and was chosen in order to achieve smooth animation and reliable recording of data). In order to synchronize the streams, the captured data was linearly interpolated. To eliminate jittering, data captured with the eye-tracker was smoothed with a small moving window.

The stimulus was presented as a single item on a black background. The colour scheme was chosen so it would not irritate the participants eyes, while the size of the dot was approximately 2° , in order to fill out the *fovea centralis* zone of vision.

In order to verify that the subject is looking at the stimulus, we introduced a feedback mechanism. At random moments, the item being tracked changed from **observed state** (represented by a green circle, Fig. 6.5b) to **unobserved state** (displayed as the yellow cross, Fig. 6.5a). The subject had the task to change the “yellow cross” into the “green circle” (desired state) whenever the change of appearance was observed, by clicking the left mouse button. The third state of the stimulus item was **tracking state** (shown as a white square, Fig. 6.5c) which was used to guide the subjects gaze during the *smooth pursuit* travels, during which the subject had no obligation to perform any action (except for the implicit task to follow the item with the gaze). All states and transitions are shown in Figure 6.6.

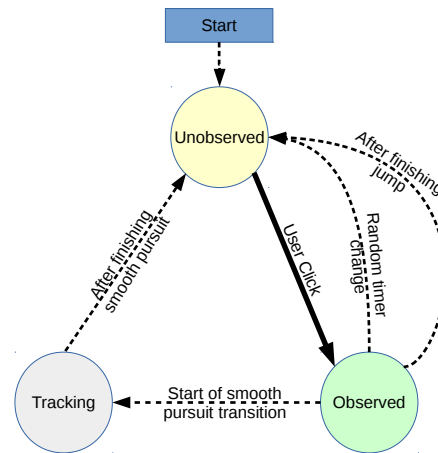


Figure 6.6 – State transitions used in the controlled experiment. Dashed lines represent automatic transitions, and full line represents the effect of the user's input.

Based on the subjects clicking, we took into consideration only the periods when the stimulus was in the *observed state* - which indicated that the subject was looking at the stimulus. Our procedure switched the item's state to *unobserved* whenever it changed its position (i.e. we assume by default that the subject does not see the stimulus). The switch to *unobserved state* was also triggered by a random number generator, which switched the state to unobserved every couple of seconds (on average). We recorded the state of the stimulus with the same resolution as the ground-truth position (10fps). For each subject we also counted the number of wrong clicks (subject clicked when the object was in the *observed state*) and the duration of time the stimulus was in the *unobserved state*.

6.2.2.2 Capturing procedure and instructions

The controlled experiment was conducted on 30 subjects from the student and employee population of EPFL, 5 of which were rejected due to bad quality of measurements (total number of analysed participants is 25). Population had a mixed composition of age and gender, majority being male and in their bachelor studies. Only 1 student had eye-sight problems, and was checked before the experiment if he was able to see the stimulus well.

At the start of the experiment, the subject was seated on a fixed chair located at the previously measured distance from the projection surface. The calibration steps included determining the centre of the subject's field of view (focused on finding the horizontal and vertical offset at which we should project the stimulus), and the 3-point calibration of the eye-tracker.

The instructions to the subject were:

- to behave naturally, that there are no constraints on the movement of head or eyes;
- that there will be a single item displayed on the screen and that it will move only along the horizontal axis (to avoid confusions and searching for the stimulus up or down) and that the subject should focus on the item for the entire duration of the experiment (we avoided saying “follow the item with your gaze” because the phrase lead to confusion, with some subjects thinking they were not supposed to move their heads);
- that there are 3 appearances of the item (explained only by appearance, not by their role), that the change from the “green circle” to “yellow cross” will occur randomly, and that the subject’s goal is to keep the item in the “green circle” state as much as possible by clicking the mouse button (which will change the state of the item);
- that a single mouse click is enough, that there are no needs to position the mouse on the location of the item, and that random clicking is recorded and penalized;
- that the user should click to change the items’ state only if (s)he is looking at the item.

The participants did not receive the explanation about what is being measures, and they were focused on the quickness of response and the correctness of feedback (mouse clicks and looking at the item). Each participant received a monetary reward for performing the experiment independent of performance.

6.3 Research questions

In order to create a reliable model of human gaze for our head-tracking experiment we needed to investigate a number of properties. Our tests are aimed at collecting additional information about human gaze and re-evaluating existing assumptions on a purposefully created dataset.

1. **Usage of the visual range** - Even though the horizontal limits of the visual range have been explored, we are interested in finding the usage patterns within the limits for our classroom scenario.
2. **Imposed recording limitations on the visual range** - For the validity of the tests it is also important to see how often were the limits of the field of view reached, and whether the imposed limit of the eye-tracker (approximately $\pm 30^\circ$) is a concern.
3. **Relationship between head pose and gaze direction** - Our main question for our experiment is to evaluate the relationship between the head pose (α_H) and overall gaze direction (α_G). We will try to determine to what extent does the head pose reflect the direction of the gaze. We aim to validate the relationship between the head pose and gaze previously found in other publications, and to model the reverse: predicting gaze direction based on the observed head pose.

4. **Head pose change over time** - We will try to see how the pose of the head changes over time we spend looking at a specific direction. Given that the stimulus is adjusted for easy finding, we expect that the initial head/eyes pose will always be reached in less than 5 seconds (minimal duration of stay for the stimulus), and we are interested how and if it will change with prolonged stay periods. Even though the main purpose of *jump* and *smooth pursuit* conditions is to model different real-world behaviours we will also test whether there is a difference between the “travel” conditions.

6.4 Results

6.4.1 Real-world usage

In the part of the experiment with unscripted behaviour, we used six recordings of two teachers who participated in our studies.

Question 6.3.1 The heat-maps of teacher's vision captured from the glasses confirmed a 2D Gaussian distribution of gaze over time under normal usage (samples of data shown in Figure 6.8). Our findings are given in Table 6.3, showing percentage of recorded samples with the associated angular limit (e.g. 25% of samples in Recording #1 were within the 9.6° of the centre of the field of view). Please note that these findings are based only on the position of the eyes, and do not record the orientation of the subject's head.

Percentile	25%	50%	75%	90%	95%
Recording #1	9.6°	12.75°	16.58°	20.8°	23.5°
Recording #2	6.6°	10.04°	15.02°	20.82°	24.05°
Recording #3	5.1°	8.7°	13.71°	19.11°	22.57°
Recording #4	4.22°	6.97°	10.58°	15.04°	18.52°
Recording #5	3.7°	6.64°	10.71°	15.43°	18.79°
Recording #6	6.94°	10.51°	15.58°	20.47°	23.26°
Mean values	6.02°	9.26°	13.69°	18.61°	21.78°

Table 6.3 – Upper angular limits for percentiles of accumulated samples of teacher's gaze. We show angular values for each recording and mean angular limits. Visualization of data is shown in Figure 6.8.

Question 6.3.2 The major observation is that even if we take the broadest limit, 95% of the time the angle of gaze position is below 20° which is well below the previously established limits of the eye movement range. All subjects except for the last one displayed the similar pattern shown in Figure 6.8a, while the outlier is shown in Figure 6.8b. After reviewing the video footage, we noticed that the teacher in that situation had problems with the laptop and was focused for some duration of time on fixing the hardware problem which would explain the points in the lower central area.

It is important to note that the limit of the visible field recorded by the eye-tracker is approxi-

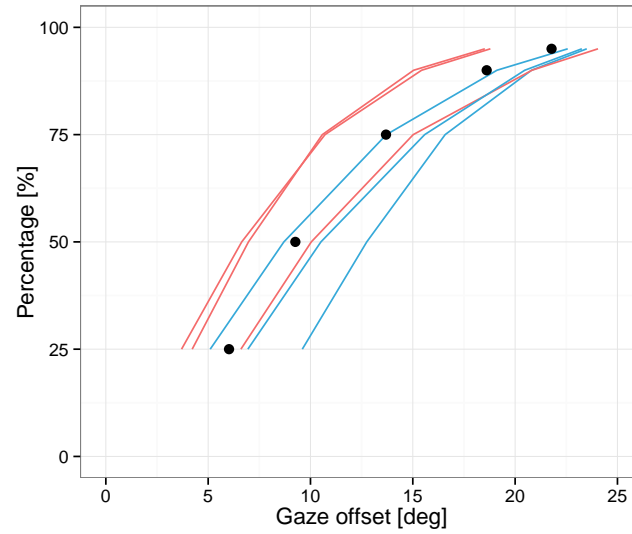


Figure 6.7 – Unscripted gaze behaviour captured from the teacher's in class. The line colours represent. subjects 1 and 2, and the black dots show the mean values for the 25%, 50%, 75%, 90% and 95% of observed gaze locations. Exact values are shown in Table 6.3.

Angle	0°	5°	10°	15°	20°	25°	30°	35°	40°	45°
+	9927	1992	3814	896	1941	3651	823	1652	3236	654
-	-	3769	1950	897	3763	1825	846	3404	1239	668

Table 6.4 – Number of samples collected for the positive and negative values of each horizontal angle.

mately $\pm 30^\circ$ horizontally, which shows that the limits found here were not imposed by the equipment used for the experiment.

6.4.2 Controlled experiment

Depending on the duration of the stay we collected a varying number of samples for different angular positions. The number of samples collected is given in Table 6.4. Other factors which influenced the number of samples were *i*) the speed at which the student reacted to the stimulus switching to an *unobserved* state and *ii*) the quality of the eye-tracking.

With the information about the horizontal head angle (α_H), angle of the eyes (α_E) and ground-truth (α_{GT}), we were able to calculate the errors for each horizontal angle $E(\alpha_{GT})$ using the formula:

$$E(\alpha_{GT}) = |\alpha_{GT} - (\alpha_E + \alpha_H)| \quad (6.2)$$

Obtained data is shown in Figure 6.9. Even though many outliers are present, the mean values

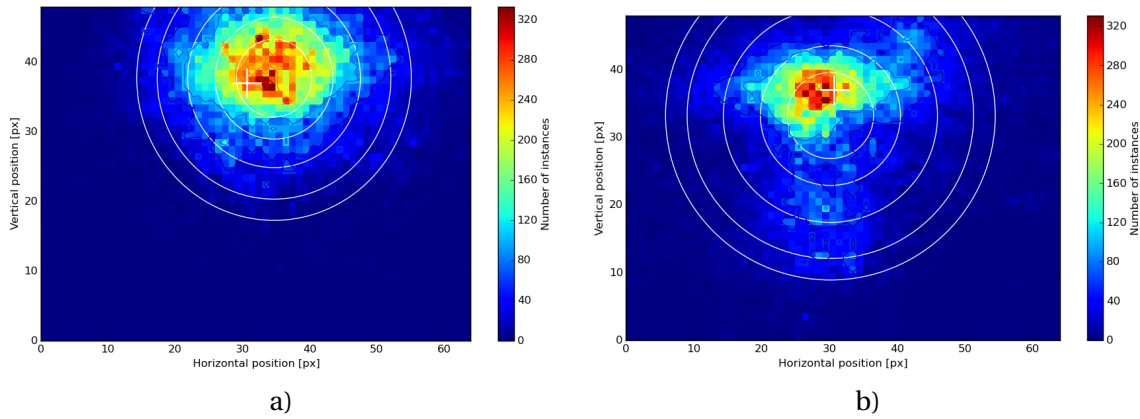


Figure 6.8 – **a)** Heatmap of a single recorded teacher. Note that the centre of the distribution is almost identical to the centre of vision (marked with a white cross-hair). White circles designate (from the centre outwards) the 25%, 50%, 75%, 90% and 95% of all captured gaze locations. Binning factor for the heat-plot were zones of 20x20 pixels. **b)** Deviating sample of teacher's vision.

of errors for all angles are below 2.5° .

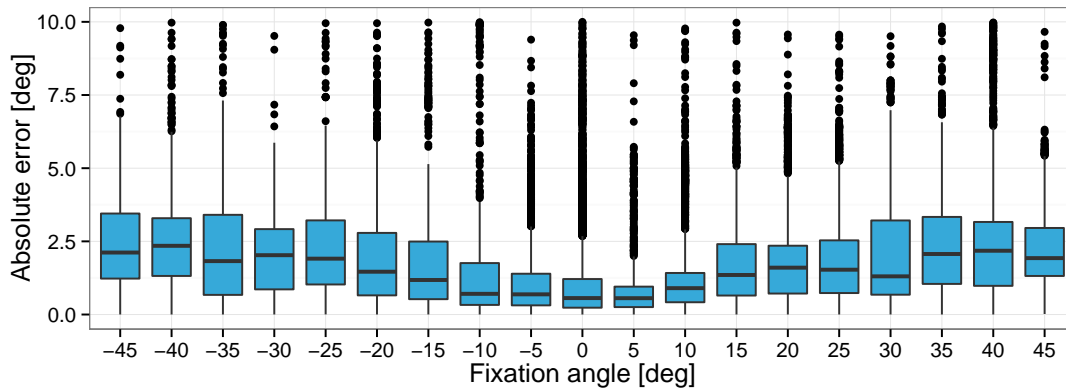


Figure 6.9 – Absolute errors in angles for each of the fixation locations. Outliers of greater magnitude were present, but were cut from the graph for clarity of the main body of data.

6.4.2.1 Head participation

Question 6.3.3 For modelling head participation with a single function, we used the head amplitude (absolute value of the angle of horizontal head rotation) from all valid samples. By fitting a linear model on the relationship between the absolute values of head rotation and absolute values of gaze direction we reached the equation

$$\alpha_H = 0.5601\alpha_G + 0.9215 \quad (6.3)$$

where the previously defined $\kappa_\alpha = 0.5601$. Though present and significant ($p < 0.05$), intercept in the model is very small ($< 1^\circ$) and can be attributed to the noise gathered when the head was placed in the neutral position (0°). The fitted line and the variability of the data can be observed in Figure 6.10. Our model is therefore still within the range of the previous assumptions (Ba and Odobez (2009) proposed range for $\kappa_\alpha \in [0.5, 0.8]$), but close to the lower limit. We also saw no support for the ϵ_α threshold under which the head participation is 0, because we see head contributions even in the 5° and 10° conditions.

Quadratic fit on the same data is slightly better, but the strength of the quadratic member is very small (0.007) at the cost of the linear factor (0.2734) and stronger intercept (2.2133), which we do not see as a good enough values to support the idea of head motion being a quadratic function of the gaze direction.

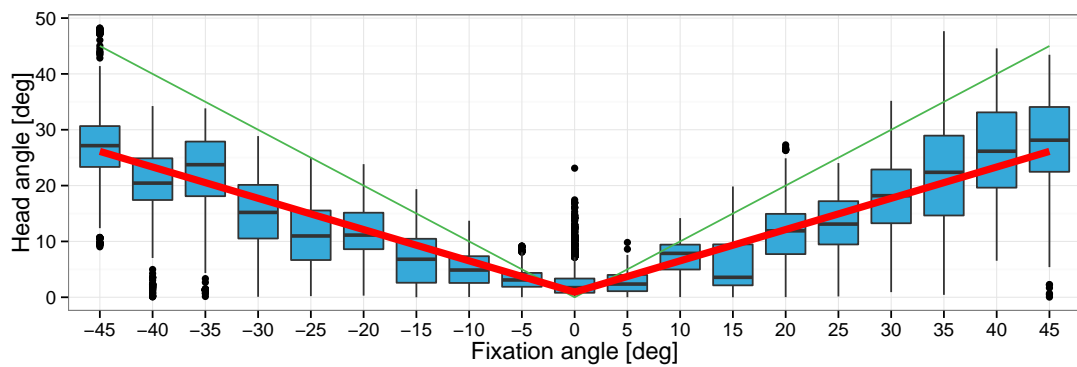


Figure 6.10 – Visualized level of head participation. Red line represents the amplitude (absolute value) of head rotation over the sampled range of horizontal angles modelled by Equation 6.3. As a reference, green line visualizes the hypothetical 100% of head participation. Box-plots show the variance of collected head rotations samples (boxes show the 25-75 percentiles, and black lines show the mean values).

6.4.2.2 Head participation over time

Question 6.3.4 In order to test how the head participation changes with different durations of stay, we analysed each fixation angle separately. The moment when the stimulus reached the horizontal angle is represented as the 0th second, and the head pose progression was tracked from the moment the user reported the stimulus as observed (usually, the time needed was less than 0.5 sec, the offsets visible in the graphs shown in Figure 6.13). Similarly, to neutralize the fixation angle amplitude and orientation, we took the initial head angle (at the 0th second) as the offset value, and tracked the changes relative to that. Positive changes represent increases in head angle amplitude (moving the head away from the centre), and negative changes reflect moving the head closer to the centre (neutral position). Graphs for all angles are shown in Figures 6.13, 6.14 and 6.15 shown at the end of this chapter.

With the exception of $\pm 45^\circ$ extremes, which were not visible for some participants (because

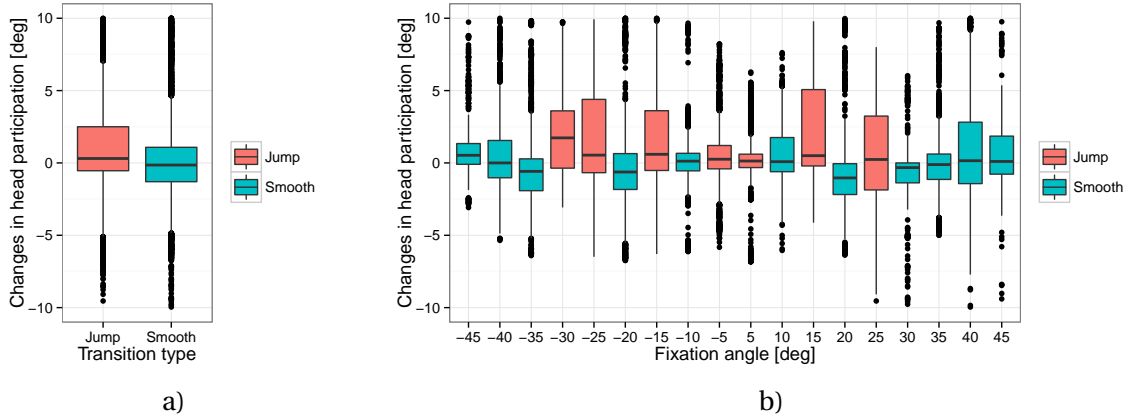


Figure 6.11 – Changes in head pose relative to the initial fixation. **a)** Shown for the two transition conditions aggregated. **b)** Changes in head pose shown for each angle, transition conditions represented with colours.

the calibration step shifted them off the projection surface), all participants observed all angles. Even though oscillations exist for individuals, the mean fitted curves remain flat for the majority of the angles. After displaying the aggregated changes per angle in Figure 6.11, we can see that the majority of the changes does not exceed 5°, and that the “jump” condition (for angles greater than the ±5°) has the more varied changes with more points shifted towards the positive. This difference was confirmed as statistically significant by one-way ANOVA, $F(1,37018) = 222$ ($p < .0001$).

Our interpretation is that in the *jump* condition, the eye-movement is dominant when seeking the target, followed by the slower head adjustment. This would explain the positive changes in head pose, with the head slowly assuming a more comfortable position. It is interesting to note that the angles where this was observed are still within the used visual range (observed in the Section 6.4.1), but participants still chose to adjust the pose.

6.5 Gaze limits from head pose

Disregarding the intercept for its minimal influence and reversing the Equation 6.3, we can model the gaze direction limits as

$$\alpha_G = \frac{1}{0.5601} \alpha_H = 1.7853 \alpha_H \approx 178\% \cdot \alpha_H \quad (6.4)$$

Given that the maximum gaze angle is 78.53% bigger than the head rotation, we took the same relationship for the minimum gaze angle as the -78.53% of the head rotation (21.47%). With that assumption, we can model the limits of the gaze as

$$\begin{aligned} \alpha_{Gmax} &= 1.7853 \alpha_H \\ \alpha_{Gmin} &= 0.2147 \alpha_H \end{aligned} \quad (6.5)$$

The main aspect which is preserved is the linear relationship of the gaze and the head, showing that at larger head angles, the gaze also becomes less predictable. In the case of upper limit, the gaze is following the observed relationship with the head orientation, and the lower limit models the symmetrical behaviour towards the neutral position. The linear nature of the gaze limits is altered by two factors:

- in case of small head angles, the gaze uncertainty does not go to 0. For this reason, we are setting the minimal gaze limits at $\pm 18.61^\circ$, observed in the unscripted behaviour (Section 6.4.1);
- in case of extreme values, the gaze direction will not go above the physical limits of the human body. We assume the maximum limit of the gaze to be $\pm 130^\circ$ which models maximum head rotation (90°) and gaze rotation (40°).

Each component listed is visualized separately in Figure 6.12a. The final function modelling the gaze limits based on horizontal head rotation is shown in Figure 6.12b.

In the next chapter we will test a number of models of represent gaze behaviour inside the limits. Our model can still be wrong for instances of sudden eye movement, but we can expect that large angles of eye rotation will be used only temporary and that head orientation becomes more indicative of gaze direction over long periods of time.

6.6 Conclusion

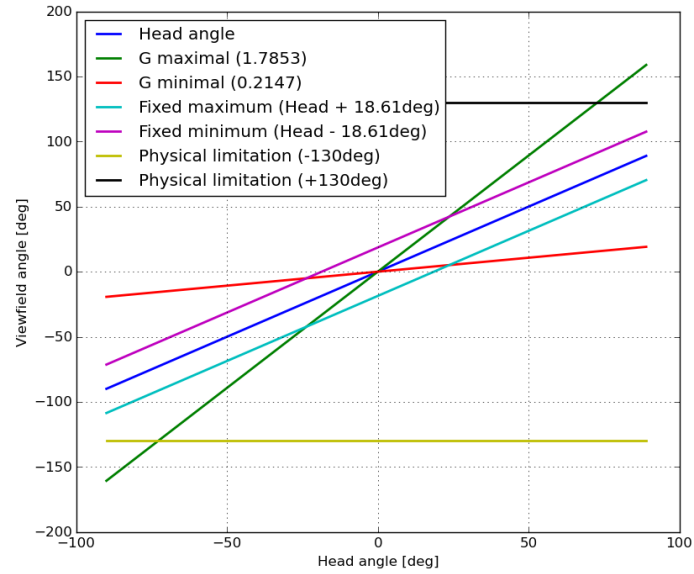
We have reached a number of conclusions with our experiments. We confirmed previous hypothesis (Ba and Odobez, 2009) that head participation can be modelled with a linear function, and found that in our sample the value of the parameter is $\kappa_\alpha = 0.5601$. Combining the linear model with additional observations about gaze behaviour, we formulated our model for estimating gaze limits from the observed head orientation.

As for behaviour of head pose over time, we observed that in the *jump* condition there is a bigger adjustment of head pose over time, with the head participating more and diminishing eye participation. Our maximum duration of the angular fixation was 20 seconds, which was not enough to simulate long fixation of the audience, but we can extrapolate from the observed behaviour that the head participation over long periods would become more prominent.

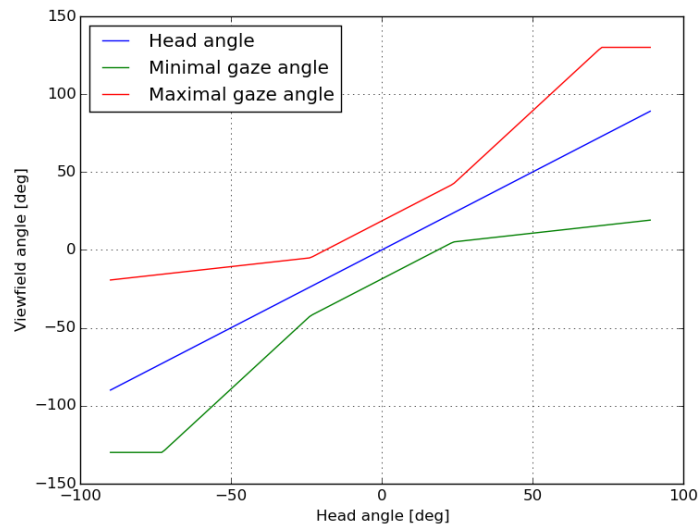
Even though the horizontal angular range of eyes motion is substantial, in our recording of the unscripted eye behaviour we observed that the angle of the eyes remains below 18.61° for 90% of the observed fixations. This further supports the assumption that the eyes, even though flexible, are not comfortably positioned at a large angle for prolonged periods of time. The eye usage can be described with a 2D normal distribution, with the centre in middle of the field of view. For our purposes the vertical motion of the eyes will be disregarded, and we

will continue by modelling the gaze as a normal distribution in the horizontal space, with the centre aligned with the head's orientation.

The experiment naturally has its limitations. In the controlled settings, the fixation periods were not long enough to see if the relationship between the head pose and eyes would evolve further. Also, the technical limitations of testing only the $\pm 45^\circ$ range force us to extrapolate our conclusions to bigger angles. This can be justified with the idea that the angles above 45° are less represented in the classroom, and that the main part of the range is covered. Finally, our setup did not provoke the subjects into “contradicting” configurations of eyes and head (e.g. head oriented left, eyes looking right).



a)



b)

Figure 6.12 – Field of view limits modelled from the head orientation. **a)** Shows all the components of the final function overlayed. Linear relationship from the Equation 6.4 is modified with imposing the minimal uncertainty of $\pm 18.61^\circ$ and limiting the extreme values to $\pm 130^\circ$ modelling the extreme head and eye rotation. **b)** The final function visualized.

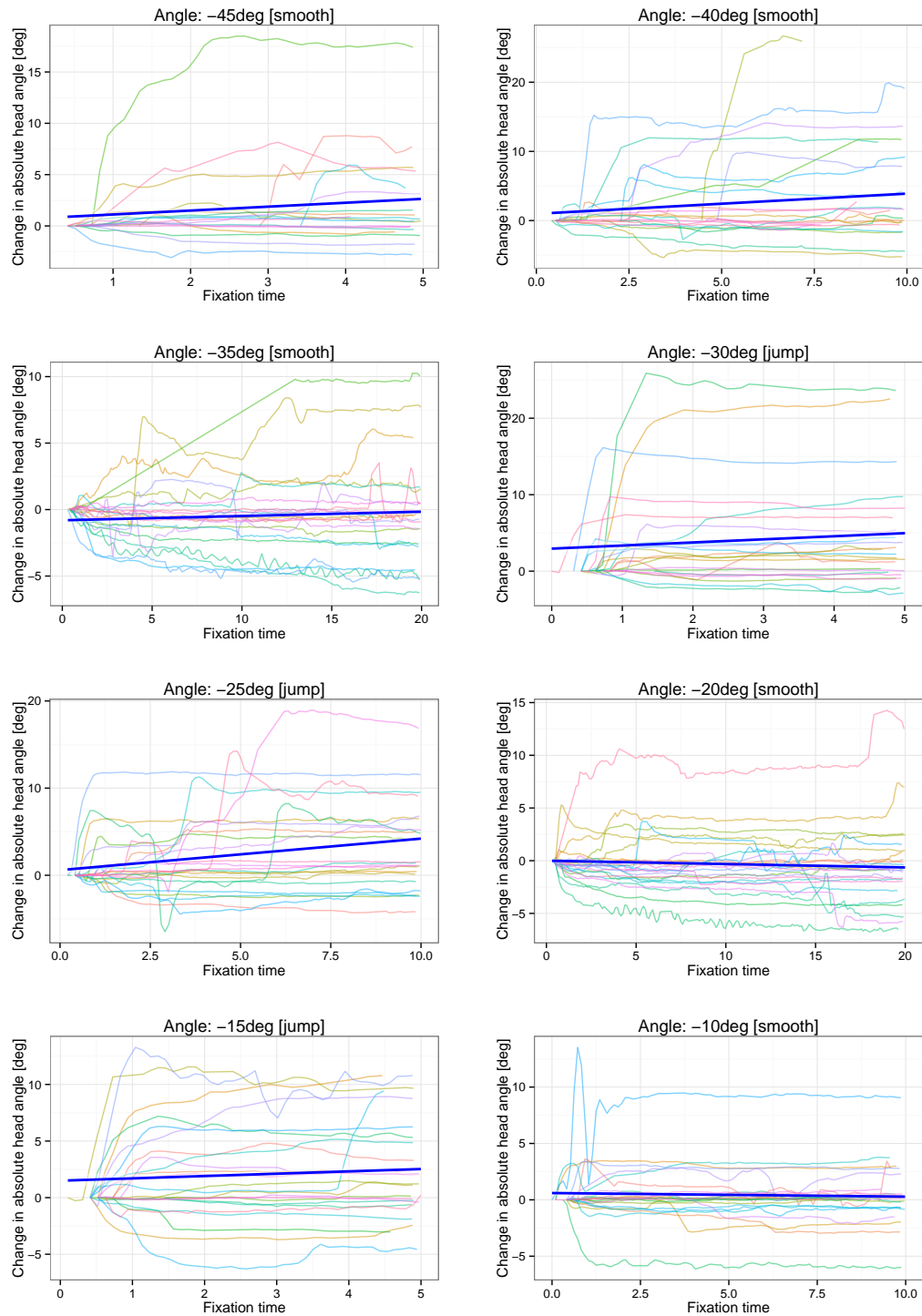


Figure 6.13 – Progression of head-pose in absolute values (positive is away from the centre, negative values are towards the centre) for different fixation angles. Travel condition and angles are indicated in the title of the graph. The angle was set to zero on the moment the stimulus was first indicated as observed, and the relative progression of head pose was tracked over time from there. Prominent blue line indicates fitted linear model on all data points observed.

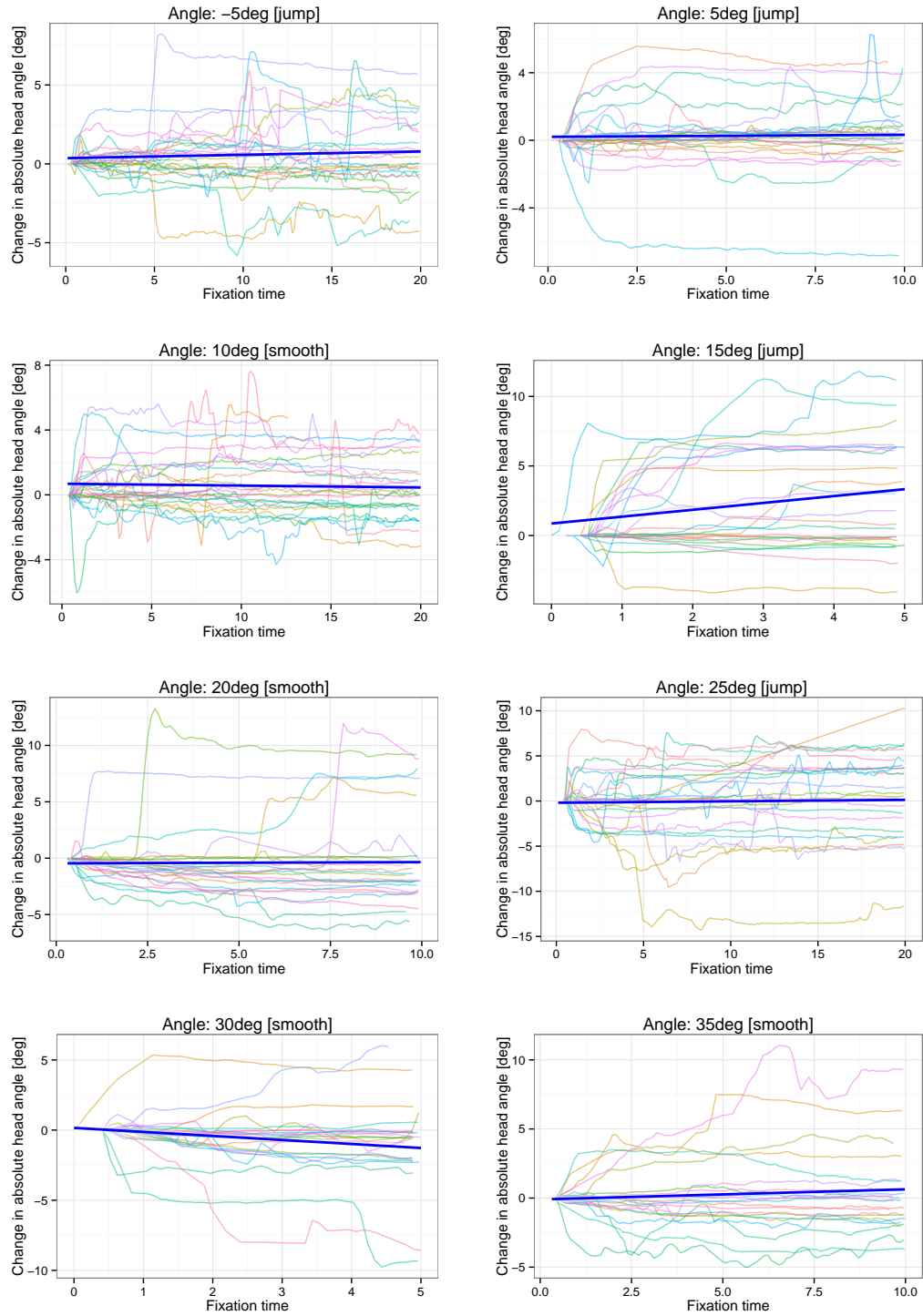


Figure 6.14 – Progression of head-pose in absolute values (positive is away from the centre, negative values are towards the centre) for different fixation angles. Travel condition and angles are indicated in the title of the graph. The angle was set to zero on the moment the stimulus was first indicated as observed, and the relative progression of head pose was tracked over time from there. Prominent blue line indicates fitted linear model on all data points observed.

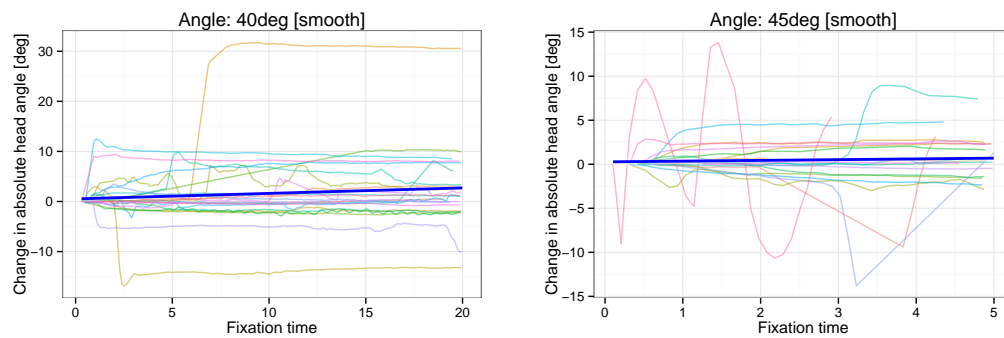


Figure 6.15 – Progression of head-pose in absolute values (positive is away from the centre, negative values are towards the centre) for different fixation angles. Travel condition and angles are indicated in the title of the graph. The angle was set to zero on the moment the stimulus was first indicated as observed, and the relative progression of head pose was tracked over time from there. Prominent blue line indicates fitted linear model on all data points observed.

7 Gaze in the classroom

Parts of this chapter have been previously published in Raca and Dillenbourg (2014) and Raca et al. (2015).

“**D**IRECT your attention” in everyday speech is almost synonymous with head turning and gaze direction. Formally explained with *top-down sensitivity control* (Section 2.1), directional properties of gaze give us the unique opportunity to use our physical adjustments as a social signal.

In this chapter we will consider head orientation as a form of direct synchronization (Section 3.1.4). As we listed in our theoretical discussion (Section 3.1.1), there is a number of sources of information in the classroom which can occupy students’ attention. Our choice of social signals was again guided by:

- spontaneity or unintentional emission of the signal,
- observability from our far-field camera system,
- characteristics of the scenario (classroom format and lecturing activity).

With that in mind, we chose to focus on the connection between students and lecturer. Our research represents complementary work to that of Stiefelhagen et al. (2006) who focused on estimating the gaze direction of a moving presenter. Similarly to the earlier research, we will also use the information about the location and movement of the teacher during the lecture. However, we will not try to estimate the exact point of teacher’s visual attention, due to the technical limitations.

Similar to Chapter 5, we will go into the technical details of our re-training procedure for Part-Based Model (Zhu and Ramanan, 2012) and additional steps needed to extract head-pose information from the audience video. The automatically-extracted features will then be analysed in relationship to attention information acquired from the questionnaires.

Our conclusions will be based on the larger data sample acquired from Populations #3 (4 recordings, 27.5 students on average) and #4 (3 recordings, 39.3 students on average). Detailed information about the sample population is presented in Section 4.2. Not all recordings will be used for all conclusions. The main limitation is the recording of the teacher's position, which is missing for some of the lectures either because of the technical problems or due to the poor tracking quality.

Our assumptions about the student activities will be the same as in the previous experiments – the students remain seated for the duration of the recording session and the camera is positioned in a fixed location. No restrictions about the format and content of the lecture have been imposed.

7.1 Data extraction

Considerable effort was put into collecting and synchronizing data. We will give a brief overview of the manually annotated data, and focus on the technical details of the algorithms used for the automatic extraction of features. All video-streams were manually synchronized prior to their processing.

7.1.1 Manually annotated data

Apart from the questionnaire data (detailed in Section 4.3) used for sampling students' attention levels throughout the lecture, there was a number of manual annotations used. Same as in our motion study, each student was annotated with a rectangular region in which they reside during the duration of the lecture/recording and assigned a unique ID (maintained over all recording sessions).

Additionally, we manually determined the **zero-angle**, which represent the angle at which the student is observed from the camera's view-point when looking straight towards the centre of the presenter's region. The angle was used as the angular offset for later head observations, in order to transform the head orientation from the camera's perspective to global coordinates. We initially tried to determine the zero-angle by geometrical means (calculating the angle from the coordinates of the student and the projection area), but the results were not satisfying (the short time for set-up of the recording units before the lectures prevented collecting the extrinsic parameters of the camera needed for precise data).

7.1.2 Head detection with Part-Based Model

There is a number of head detectors available for usage based on published research – OpenCV (Bradski, 2000) implementation of the general object detector developed by Viola and Jones (2001), or DLib's (King, 2009) object detector based on Histogram of Oriented Gradients (HoG) (Dalal and Triggs, 2005) and SVM. A broader overview of all technical approaches is given in

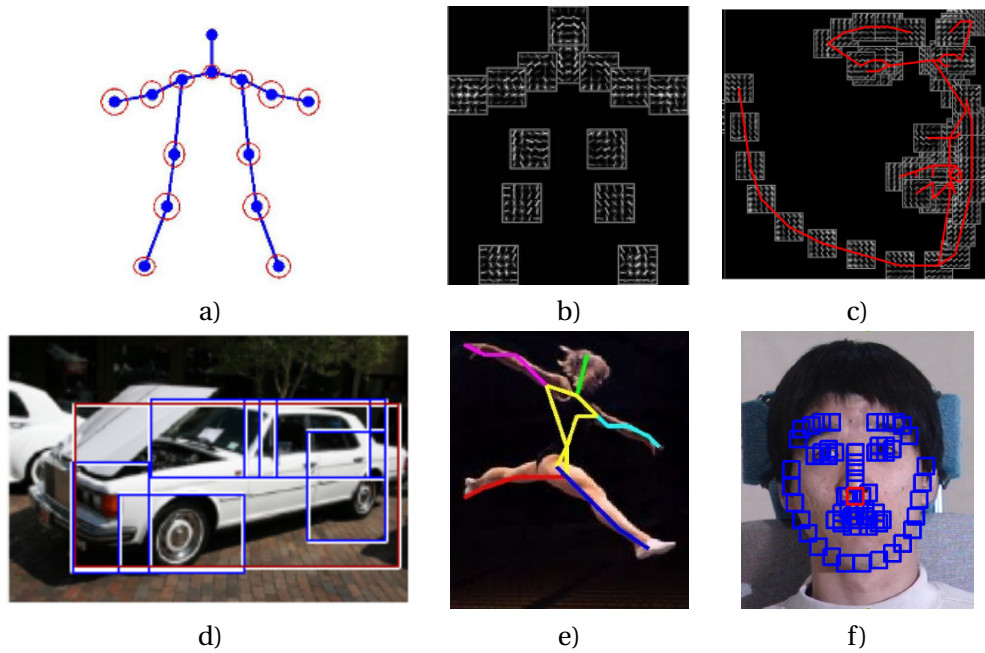


Figure 7.1 – Part-based models **a)** Visualization of the tree-of-parts for the detection of human figure; **b)** HoG signatures for each of the parts visualized; **c)** model for face detection; **d)** example of DPM for car detection; **e)** example of pose detection/estimation; **f)** example of face detection with visualized sub-parts. Images taken from (Felzenszwalb et al., 2010; Yang and Ramanan, 2011; Zhu and Ramanan, 2012).

Zhao et al. (2003).

Even though most detectors can claim decent performance, their focus remains on handling frontal face orientations. Effective range for some of them can be extended by creating a detector array, e.g. by adding additional detectors for side-face as a distinctively different pose.

As our main goal is to estimate the pose, our detector needed to be flexible and cover a wide range of poses. For this reason we choose the Part-Based Model (**PBM**) presented in the work of Zhu and Ramanan (2012). Apart from setting state-of-the-art results at the moment of publication, PBM provided the needed flexibility in detection and a formulation that solved three problems in parallel – face detection, pose estimation and facial landmark localization.

PBM is based on earlier research of pictorial structures (Fischler and Elschlager, 1973), which got popularized in the formulation of Deformable Parts Models (**DPMs**) done by Felzenszwalb et al. (2010). Both approaches try to handle the intra-class visual variability (e.g. smiling and crying face) by modelling non-rigid deformations with their models. The underlying idea is that the targeted object consists of sub-parts with fixed appearance (eyes, nose, chin) and that the variability comes from a spatial deformation between these parts (contracting, extending, rotating).

Both DPM and PBM use Histogram of Oriented Gradient (**HoG**) (Dalal and Triggs, 2005) for representing sub-parts. A trained HoG can be viewed as a 2D average of orientations of edges

in the image (shown in Figure 7.1b). This enables the part description to capture the general shape pattern, while ignoring changes such as colour and small lightning changes. The score of HoG is calculated in a “sliding window” manner, in which the 2D filter of size (w, h) is applied to every image location (x, y) to calculate a score whether that location fits the pattern described by the HoG.

Sub-parts of the object, described by HoG filters, are organized in different structures. While DPM used a “star” model (position of each sub-part is relative to the position of the main “root” node), PBM replaced it with a tree-like structure (each sub-part has a single parent sub-part with no loops in the connection graph describing the whole model). In both cases, the deformations were described as a quadratic function which models the location of the sub-part relative to its parent node. The parent-child relationships are shown in Figure 7.1a as lines between nodes which represent the sub-parts (depicted is the tree structure of PBM).

The final score (S) for the PBM model for image I and configuration of sub-parts L can be described with formula

$$S(I, L) = \text{App}(I, L) + \text{Shape}(L) \quad (7.1)$$

where App measures the total appearance similarity of the image with our HoG descriptors and Shape measures the likelihood of that specific configuration of sub-parts.

Not all appearance changes can be modelled with re-arrangement of parts (e.g. a car viewed from a side-view looks very differently than a car viewed from front - still it represents the same class of objects). For this reason, the concept of **mixtures** was introduced. Mixtures are independently trained deformation and appearance models. In the final model, mixtures act as an array of detectors, where the best scoring mixture and associated sub-parts configuration wins. For each appearance sub-class training set can be split manually or automatically based on a heuristically defined criterion (e.g. width of the bounding box, used in Felzenszwalb et al. (2010), or different steps of horizontal rotation of the head). Example of PBM mixtures modelling horizontal head rotations are given in Figure 7.2.

7.1.2.1 Part-based model reformulation

Our main difference from the initial part-based model is our focus on lower resolution images (given that typical face in the recording is approximately 50x50 pixels, and the minimal recommended size of the original model is 80x80 pixels). In order to achieve this, we simplified the PBM model so that individual mixtures consist of 8 to 11 sub-parts (in the side view and frontal view respectively). We also normalized the training image size to 50x50 pixels, which represents the approximate size of our detections.

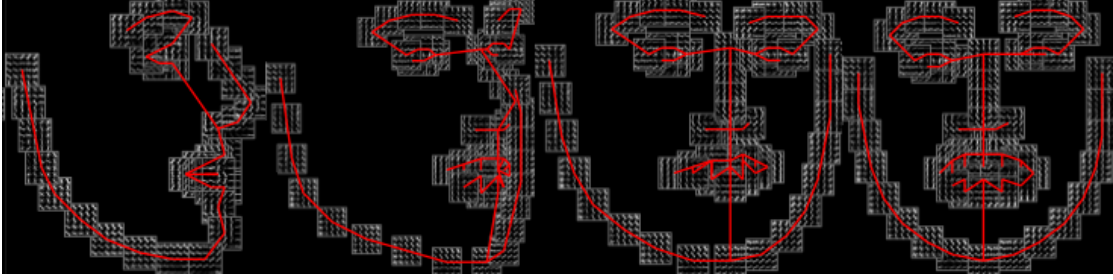


Figure 7.2 – Mixtures of the original part-based model for detecting faces. A total of 13 mixtures was used by Zhu and Ramanan (2012) for modelling different yaw angles, acquired from the MultiPIE dataset (Gross et al., 2010).

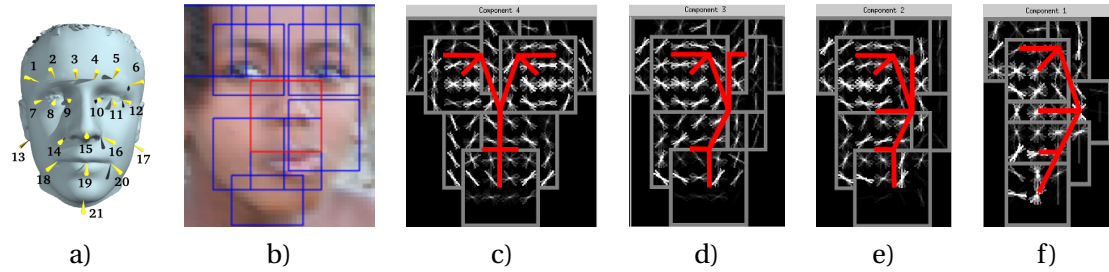


Figure 7.3 – **a)** Facial landmarks of the AFLW dataset. We used only a subset of all features. **b)** Example of the training image with correctly placed detections **c-f)** First four mixtures of the final 7-mixture model. Red lines represent the deformation-tree structures of the facial model, with the HoG visualizations of each part in the background. Angles represented are 0° , 30° , 50° and 70° of yaw. Images **a**, **b** are copyright of (Koestinger et al., 2011).

We used the *Annotated Facial Landmarks in the Wild (AFLW)* dataset (Koestinger et al., 2011) for training. The dataset provided us with both variety of head poses and annotated facial landmarks needed for training the deformation model. Given our task, we focused on the variety of horizontal (yaw) angles, and eliminated examples with extreme roll angles ($> \pm 60^\circ$). In addition, all pictures were mirrored to increase the training set for angles of same amplitude and different sign.

We also included the 1280 negative samples from the *INRIA Person* dataset (Dalal and Triggs, 2005). Pictures mostly consist of urban and architectural images without people.

Similarly to the training procedure for the original PBM model, which trained 13 mixtures based on horizontal angles in 15° steps, we trained 7 mixtures dividing the available yaw range into equal sections (step size 21° , frontal face group was merged from 2 groups to increase the number of training samples). The mixture step was chosen as a compromise between the available number of samples available for each mixture, with the emphasize on training the frontal face. Details are given in Table 7.1. When estimating the head orientation from the winning mixture we used the middle of the training range of angles used (see “Mixture designation” and “Angle range” in Table 7.1).

Each detection instance consists of:

1. detection score,
2. winning mixture number, which also represents the initial estimation of the head's yaw angle ("mixture designation"),
3. list of mixture's sub-parts (facial landmarks), represented as rectangles around the location of each sub-part in the image,
4. bounding box - a rectangle around all facial landmarks, representing the boundary of the face. The surface of the bounding box was used in all measures of overlaps between detection and another surface (e.g. the student's annotated region).

For easier visualization, we will display detections as dots in some images of this chapter. Each dot represents the centre of the detection's bounding box.

Mixture number	1	2	3	4	5	6	7
Mixture designation (°)	-70	-50	-30	0	30	50	70
Angle range (°)	-84, -63	-63, -42	-42, -21	-21, 21	21, 42	42, 63	63, 84
Num of training samples	2115	2146	4408	6712	4408	2146	2115
Num of parts in mixture	8	9	10	11	10	9	8

Table 7.1 – Details on the training split for different mixtures (modelling different yaw angles) of AFLW dataset and PBM detector. Mixtures are visualized in Figure 7.3.

In order to test detectors coverage of the pose space on a reliable ground-truth, we used Pointing'04 (Gourier et al., 2004) dataset. The dataset of 30 subjects simulates different lightning conditions while sampling the horizontal angles from -90° to $+90^\circ$ in increments of 15° , and vertical angles -90° to $+90^\circ$ in increments of 30° . Given that it is know that a person exist in each of the test images, we additionally annotated the facial regions and judged true/false positives based on the overlap between the ground truth rectangle and the bounding box of the best-scoring detection (threshold set at $> 50\%$ of surface of the ground-truth covered with the detection's surface). Examples of test samples are shown in Figure 7.4.

We reached 56.49% correct detections over all poses. This result is raised to 64.87% if we consider only poses with pitch angles (vertical rotation) between -30° and 30° , which are more frequent in our scenario. Successful detection distribution over specific head angles is shown in Figure 7.5. With the exception of the more extreme poses (such as more than $+60^\circ$ pitch), pose space is covered reasonably well. We speculate that the failure of the detector in case of a raised head is stronger because the appearance of facial parts drastically changes in such cases (e.g. bridge of the nose and eyebrows are not visible).

Given that the detection was the most processor-demanding part of the process it was executed separately from all other steps. On our workstations, processing of a single video took around 2 days with the C++ implementation of the PBM algorithm (Bristow, 2012). Due to

these technical requirements we were unable to carry out experimentation with different methods in combination of repeated execution of the detection step. We acknowledge that this has possibly limited our formulation and that improvements could be possible with better performing detection algorithm. As we previously stated, our observations are on the PBM model as it was the most flexible detector, and because of this – best suited for our experiments.

Output of the detection step was an archive of detection instances per frame of video. In order to eliminate highly overlapping detections, we performed non-maximum suppression (NMS) on the detections. NMS algorithm finds sets of detections which overlap above the given threshold (we used >90% overlap of detections' bounding-boxes) and keeps the best scoring detection from each set. With this, our output was reduced to around 2GB of zipped CSV files per video.

7.1.3 Pose estimation

The PBM formulation allowed us to detect faces and estimate their yaw angle at the same time by using the winning mixture number. Our next goal was to refine the results by using support vector machines for regression, which can be used on top of the detector's output. This turns the joint formulation of detection and pose estimation from the PBM, into a sequential process where we use the rough information from the detection step as input to the finer horizontal angle estimation. Our training was based on the previously used AFLW dataset, with regression algorithms from the DLib library (King, 2009).

This discretization of horizontal angles introduced by the mixture training caused a yaw angle mean absolute error (MAE) of 7.07° in the AFLW training set (Table 7.2).

The mean absolute error of mixture-based yaw estimation over all horizontal and vertical of the Pointing'04 dataset is shown in Figure 7.6a, and accumulated errors for each horizontal pose (errors over vertical poses summed) are shown in Figure 7.6c.

Knowing the location of facial landmarks from the detection step, geometrical methods were the logical method for increasing precision. We experimented mainly with different feature

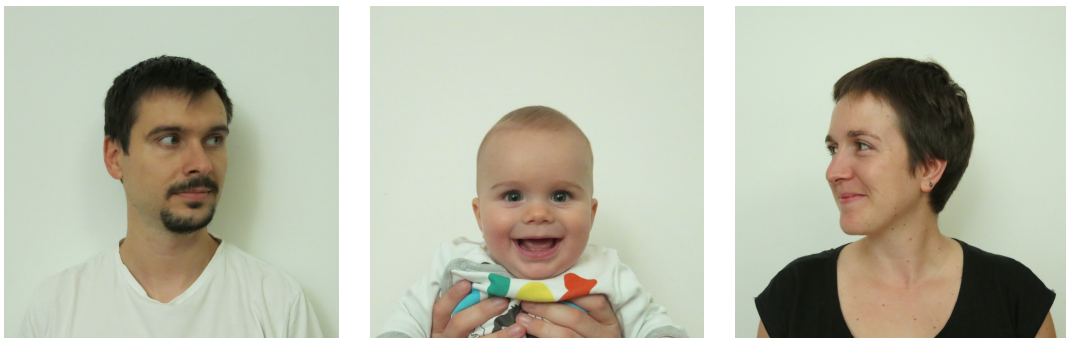


Figure 7.4 – Illustration of the format of images used from Prima Pointing'04 dataset for testing.

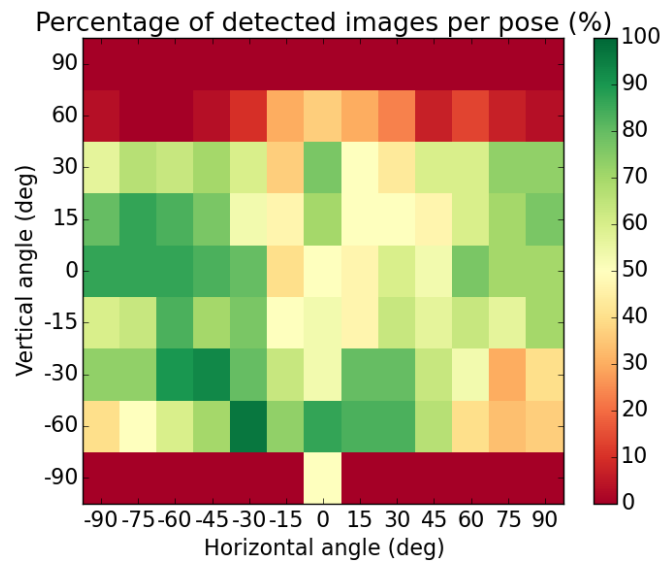


Figure 7.5 – Percent of correctly placed detections by the PBD tested on the Pointing'04 dataset. Overall detection score is 56.49%.

combinations, building on the findings reported by Murphy-Chutorian and Trivedi (2009). Training was done on approximately 20500 samples (extreme roll angles were excluded) and tested on a held-out test-set of 2200 randomly chosen images, also from AFLW. Results of this evaluation are shown in the 2nd column of Table 7.2. After testing a number of feature combinations and kernel types/parameters, our final regressor is based on a set of four features:

- normalized location (x,y) of the nose within the face bounding box – modelling the displacement of the nose depending on the yaw angle;
- width to height ratio – capturing the property faces appearing narrower when side-turned;
- winning mixture angle estimation – information captured from the detection step.

Kernel type	MAE (AFLW)	MAE (Pointing'04)
original mixture estimation	7.07°	17.56°
rbf	5.38°	18.16°
linear	6.00°	19.67°

Table 7.2 – Comparison of different regression features combinations and kernels for facial pose regression

In addition to testing models on the AFLW dataset, we re-ran the angular precision tests on 2790 images of the Pointing'04 dataset and gave the mean error in the 3rd column of Table 7.2.

We did not find a significant improvement in on the test dataset (Pointing'04) after running the SVM. Even though tests on the AFLW dataset shows improvements in angle estimation for both linear and RBF kernel (1.07° and 1.69° improvement in the mean absolute error (MAE) scores), scores on the test set showed degradation in MAE by 0.6° and 2.11° respectively. We still decided to use the model based on the RBF kernel in our results because of two reasons *i*) degradation in the results is not that severe and potentially the AFLW dataset captures the precision better with a finer coverage of the horizontal rotations range *ii*) the smoothing of transitions gives us better chances of capturing smaller changes in student's behaviour, given that the threshold of 20° for registering a shift in pose is too high.

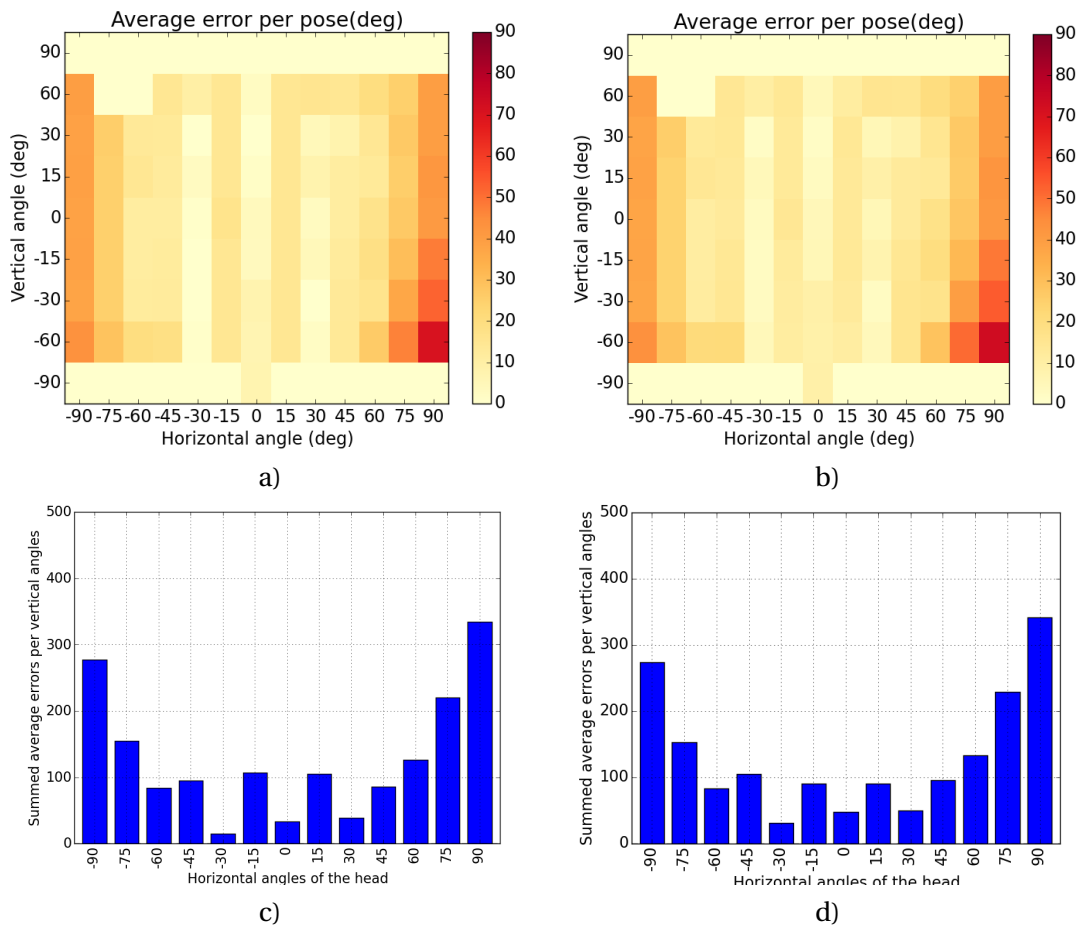
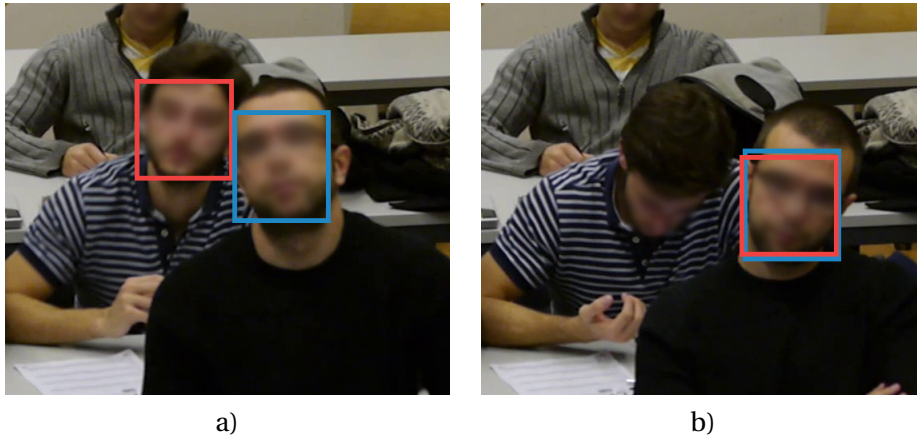


Figure 7.6 – Average angular error per head pose (horizontal and vertical angles) **a)** Original parts-based detector (mixture classification) and **b)** rbf regression based on 4 features. We accumulated the vertical error to produce the total error per yaw angle for **c)** original parts-based detector and **d)** rbf regression.



*Figure 7.7 – Problematic situation for detection assignment. **a)** Both faces are visible and correctly assigned to persons. **b)** One face is lost, with the other detection being double-assigned (shown here) or miss-assigned.*

7.1.4 Head pose tracking

After detection and angle refinement, we are facing a problem similar to the motion assignment we described in Section 5.1.4. Even though the simplest approach to the problem is to pick the best-scoring detection in the person's region as the correct one, this fails because of the overlap of student's regions – if a person's face is occluded, a neighbour's face will be picked (example shown in Figure 7.7).

We relied on the same principles of outlier-detection elimination and temporal consistency as with our motion extraction algorithms to improve the quality of our estimation about the person's behaviour during the lecture.

7.1.4.1 Outlier elimination

Unlike body motion, head locations represent a slow-changing signal. From our observations, heads remain still for long periods of time (in terms of location in the video) and typically assumes a couple of distinct positions due to body shifts during the lecture.

From those observations we can assume that the detections from the entire recording will be highly clustered at a small set of locations (shown in Figure 7.8). With this, our first step for eliminating outliers is done by pruning the detections with small number of neighbours. Because no fixed thresholds could be established, our settings were conservatively set to eliminate the 0.5% of detections with the least number of neighbours. The process was executed in four steps:

- accumulating all detections observed during the lecture,
- calculating the number of neighbouring detections for each detection within the radius

of 40 pixels,

- building a histogram of the number of neighbours from the observations in the video, and from there determining the threshold number of neighbours for rejecting 0.5% of the detections,
- pruning the detections based on the found threshold value for the number of neighbours.

Our second task was to distinguish between clusters of detections created by different persons and increase the reliability of detection assignment. Our initial attempts at solving the problem were “locally” oriented (per person’s annotated region). Several failed attempts were made to make a pose prior by using either:

- geometrical means – given that the head of the person is most likely located in the horizontal centre and vertical upper half of the student’s region. Approach failed because this was not constant over all camera viewpoints.
- fitting a 2D Gaussian distribution – using all observations to fit the most likely centre for the person. The approach failed because the overlapping regions had at some cases two or more cluster of dots which situated the centre of distribution between them.

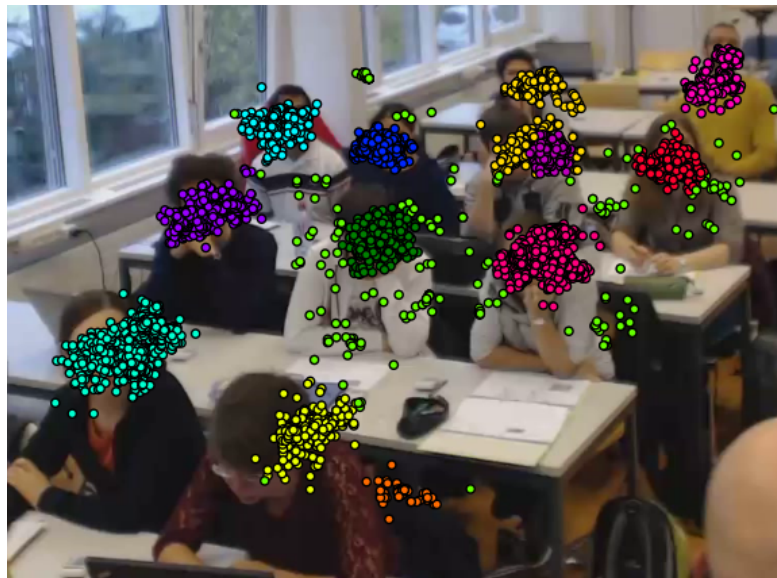


Figure 7.8 – Dense groups of detections around typical head locations (each dot represents the centre of the detection). Colours represent different GMM mixtures fitted to each cluster of detections. Beside associating the clusters with student ID’s, orange cluster in the lower centre would be marked as “outlier” and bright green cluster scattered over several students would be marked as “invalid”.

- fitting a Gaussian Mixture Model (GMM) – no general rule about the number of mixtures was found within one region and the fitted mixtures did not have a pattern to distinguish between the “positive” and “negative” clusters.

The gradual increase in complexity led us to consider the problem on a per-viewpoint bases (for the whole image captured by a single camera). We developed a semi-automated approach which we called **labelled GMM**. Instead of trying to fit a GMM on a per-region bases, we fitted one GMM on all detections observed from a single camera over the entire duration of the recording. The number of mixtures for the fit was set to $2N$, where N is the number of student regions annotated in the view (i.e. we assume an average of 2 poses for each person). GMMs were fitted using the *sci-kit learn* library (Pedregosa et al., 2011).

After the fitting step, each mixture of the GMM was manually marked as either connected to a specific student ID, invalid mixture or an mixture fitted on outliers. This enabled us to remove clusters of false detections (usually caused by a bag or a chair part) and to associate the student with a varying number of clusters, softening our assumptions about a specific number of clusters per person. The mixture fit was not ideal in all situations and in case the mixture was too wide or merged several students it was labelled as “invalid”. Next, we will describe how we merged the outlier elimination steps into a single algorithm.

7.1.4.2 Head motion

After defining our steps for eliminating outliers and associating measures with individuals, our next task was to connect observed head detections into temporally consistent sets of measurements and assign it to a student. The output of the previous steps has given us a small subset of detections with high reliability of location and pose (detection and regression steps). Because of this, our “tracking” has a simpler purpose than the regular tracking algorithms such as the *Kalman filter* (Kalman, 1960).

From the Computer Vision point of view, Kalman filter and related particle filters are based on two main components:

- **motion model** - serves as a mathematical formulation of object’s motion. Common formulation models speed and location of the object in order to predict where is the objection going to be in the next time step.
- **observation model** - represents a similarity measure between the known appearance of the object and a location in the observed image (e.g. how much does a specific section of the image look like a face).

Motion and observation model collaborate interchangeably. In simplified terms, motion model directs the search for the object based on previous knowledge of location and speed,

and observation model corrects the estimations based on the information from the current image.

In our specific case:

- Because of the discrete set of observations provided to the tracking step, we do not have a complete observational model. We have a set of detections from which we choose the most likely one.
- Our motion model expects that the detection will remain at the same location or in close proximity. Note that we are talking about the location of the head in the image, and not the yaw angle.

These circumstances allowed us to formulate a simplified tracking algorithm, which is shown in Figure 7.9. The algorithm can be divided into three sections: **i)** candidates filtering **ii)** selecting the best detection and **iii)** assignment and search radius correction.

Candidates filtering is narrowing down the list of detections considered as the next detection candidate for the current student. This is done in two stages:

- Filtering of candidates based on student's region. We set a lower limit on the overlap between the detection's bounding box and student's annotated region (more than $> 50\%$ of detection's surface needs to overlap with the student's region).
- Filtering based on the labelled GMM information. Each detection is connected with the most likely mixture of the GMM. Detections which are explained by the GMM mixtures labelled with the current student's ID are kept for further processing.

Selecting the best candidate is based on a greedy algorithm, and its main purpose is to prevent jumps between high-scoring detections within the region. The algorithm enforces the idea that the centre of next detection will be in close proximity of the centre of the last one. This is possible because the capture frame rate (typically 25 frames per second) allows us to capture normal intensity movement in small steps.

In case a detection was selected as the next detection for a student, it is removed from the pool of available detections for other students, preventing double-assignment. This greedy approach is influenced by the order in which we process student regions. For this reason, we process the student regions starting from the ones closer to the camera (based on the idea that bigger head are more likely to have higher detection scores).

Assignment and search radius. In case no suitable candidate detection was found, we record a missing measurement. The algorithm will expand its search radius for the next frame, and remain centred on the centre of last recorded detection. This mechanism models that in case of missing values, our uncertainty about the next location of the candidate grows with time.

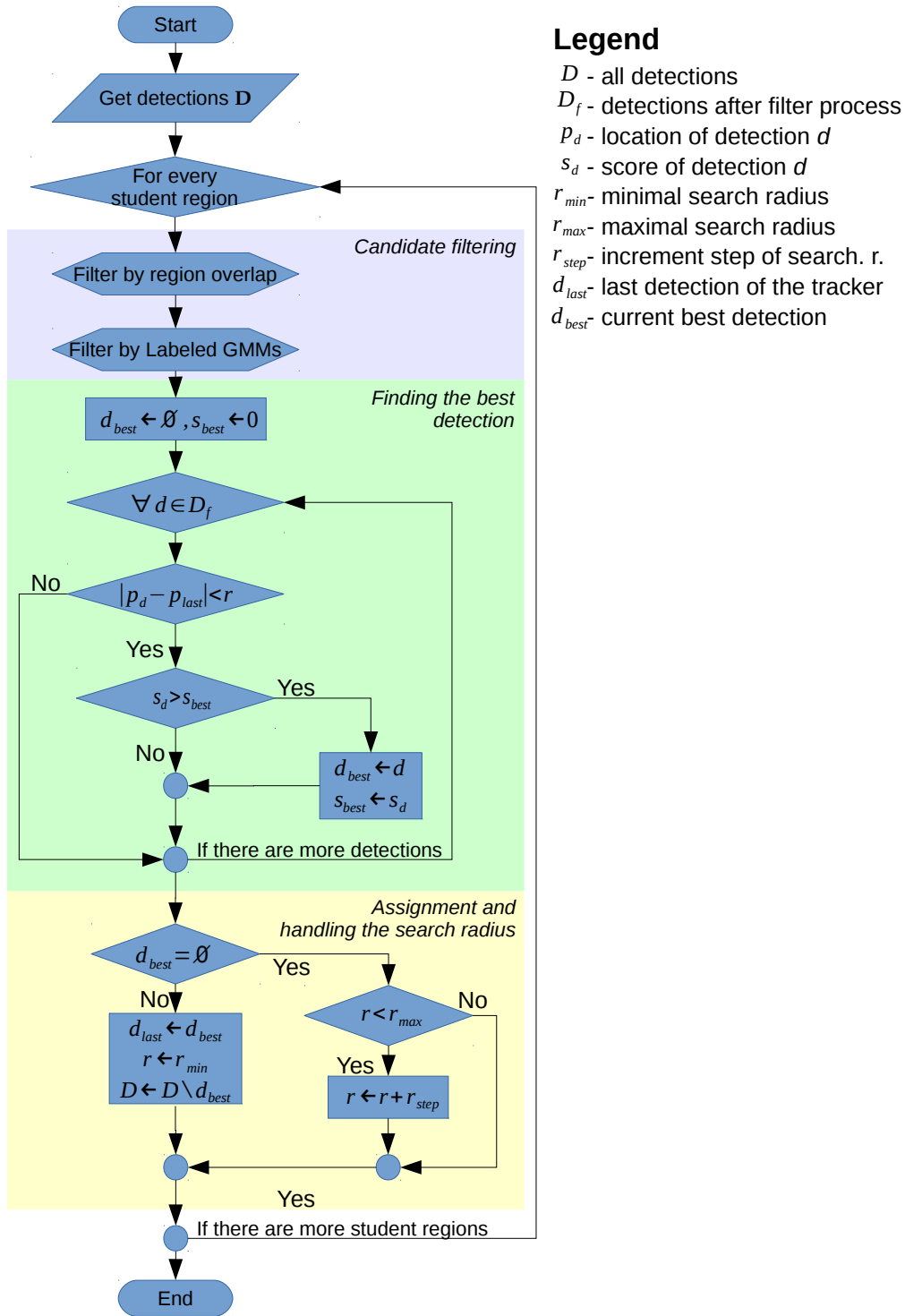


Figure 7.9 – Details of the head tracking algorithm.

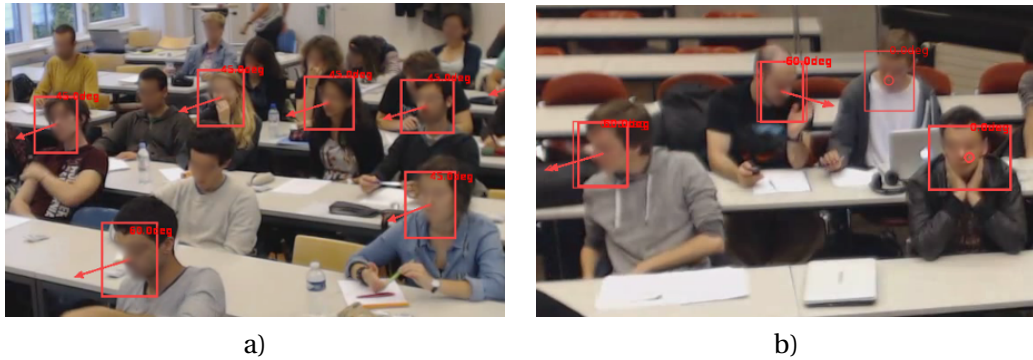


Figure 7.10 – Examples of head detections during **a)** lecture and **b)** recess.

The radius is shrunk down to the minimal value (10% of student’s region width) after each selected candidate, and will grow to the maximum of 50% of student’s region width. The speed of growth is set so the search radius will capture the maximum horizontal displacement in the region (region’s width) in 1 second.

As a final post-processing step, we did a local-window smoothing of the head angle and locations for each person to eliminate leftover noise in our measurement.

The approach was aimed at mimicking the principles observed in established algorithms such as the modelling of uncertainty and motion model, while dealing with a restricted input (preprocessed set of detections, no motion model and limited observational model). Further improvements, such as the replacement of the greedy assignment with a combinatorial optimization (e.g. Hungarian algorithm (Kuhn, 1955)) are possible. Examples of our detections are shown in Figure 7.10.

7.1.5 Teacher tracking

A separate camera, positioned in the back of the classroom, was used to record teacher’s movement. Given the low resolution of the recording, our feature set for this processing step was limited. We chose to extract the information about the location of the teacher in front of the projection area, in order to connect it to the direction of student’s gaze.

Location of the teacher was treated as a 1-D tracking problem, and only the position of the teacher along the front wall of the classroom was recorded. This was considered valid because the teachers did not leave the “teacher’s zone” (area in front of the first row of the class) in any of our recordings, and movement in the “depth” dimension of the classroom was minimal.

We used an existing *Tracking-Learning-Detection* (TLD) algorithm (Kalal et al., 2012), which relied on a manual initialization of the target. Example of video capture and tracking is shown in Figure 7.11. We focused on tracking the teacher’s head to lower the chances of occlusions. The tracking was monitored and manually re-started in case of tracking failures.

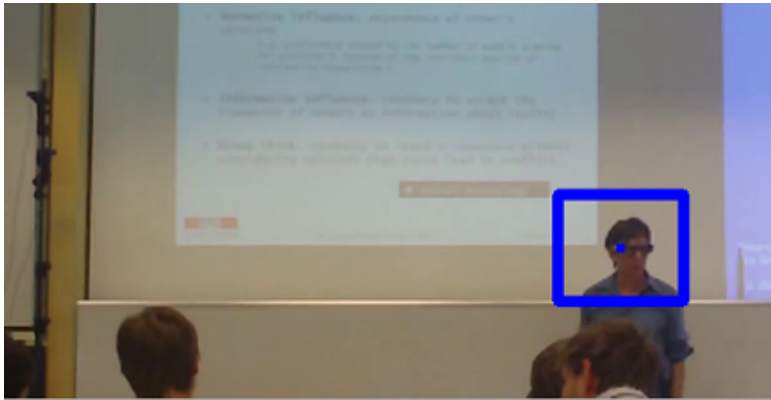


Figure 7.11 – Example of the video recording of the teacher and visualized tracking region of the TLD tracker (Kalal et al., 2012). The entire width of the front wall was captured, but the image was truncated for display.

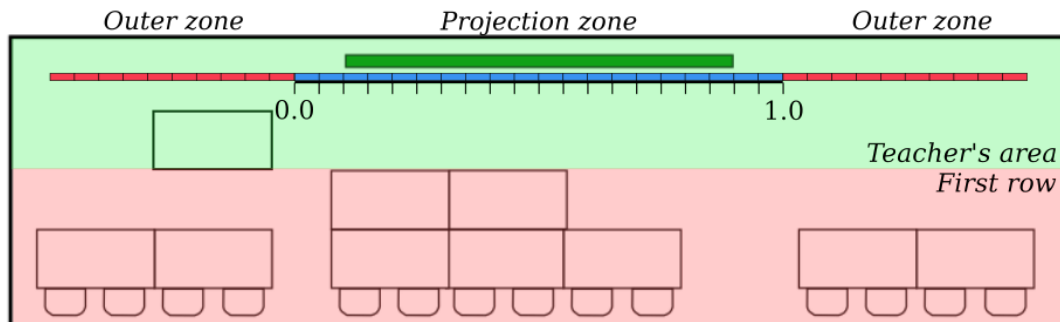


Figure 7.12 – Teacher's location coordinates. We display the 0.0 - 1.0 scale in front of the projection area, but note that the coordinates are not truncated to that range. We also display the bins used for positional histograms later shown. Bins are marked as blue and red rectangles in order to differentiate between the “projection” and “outer” zone. Bin width is equivalent to approximately half of body width.

In order to display and compare teacher's motion in relevant context, we normalized the extracted location. Values in the range 0.0 - 1.0 were used to represent the location in front of the projection area, expanded by 1 person's width on either side. We used this scale as it represented the range of locations in which the teachers were observed most of the time. Diagram of the scale is shown in Figure 7.12.

In order to compare the time teacher's spent in different locations, we discretized the locations in the teacher's area into positional bins, shown in Figure 7.12 as rectangles. We took the bin size to be approximately equal to half of body width, which we consider an minimal observable shift in position. The 20 bins in the 0.0 - 1.0 range are considered the “projection zone” and are designated with blue-coloured positional bins. Positions left and right of that range are labelled the “outer zone” of the teaching area and are shown with red coloured bin.

7.2 Defining behaviour measures

Similarly to capturing motion in Chapter 5, extracted information about head movement in the classroom needs to be formulated as a measure. In this section we define two sets of measures depending whether they are connected to an individual's behaviour or to the possible synchronization between teacher's motion and student's gaze.

7.2.1 Individual measures

All individual measurements were considered per attention period in order to associate the features to the labels acquired from the questionnaire. Because of this, each of the values describes an aspect of person's behaviour over the period of approximately 10 minutes.

Detection percentage. Assuming an equivalent quality of tracking between persons, the basic information we can extract from the detector is whether it sees a face or not. Initial assumption is that this would allow us to measure the time the student spent looking down (most probable source of true negatives) just by noting how long was the head absent.

The noise in the measurement originates from the false negatives of the detector, which is dominantly influenced by the distance from the camera. Even though we resorted to using zoom-lenses for the distant people in the class (which makes the head-size comparable to the people in the front rows), there still was a significant correlation between the row in which the student sat and percentage of time detected (Pearson's $r(192) = -0.1867$, $p = 0.009$).

In order to extract a high-level description of person's behaviour we formulated measurements which model motion and stillness of student's head activity.

Head travel records the total accumulated head travel in the horizontal direction. We ignored the potential head-travel in the periods when we did not detect the face of the student. In order to neutralize the influences of person's rhythm and distance from camera, we also defined a normalized version of the measure. **Normalized head travel** was calculated by using all the measurements of a single person per class period to determine the mean and scaled it with the variance of those measurements. Samples with less than two measurement were excluded.

We modelled the "focus" of the student with three measures connected to stillness. **Stillness** was defined as a period of time during which the angular speed of head motion is less than 10° and the overall change is less than 10° . Second condition was added in order to prevent slow, drifting movement to be classified as stillness. **Stillness period** is defined as a period of time (minimum duration of 5 seconds), in which the stillness condition is valid. Stillness periods by definition can not overlap.

From there we record for each person the **number of still periods** and the **mean duration of the still period** as the first two measures. We also included the **percentage of time spent still** as the ratio of summed duration of all still periods over the total duration of the attention

period.

Feature name	Description	Samples
Period	Period of the class (1–4), associated with the attention	776
Distance	Distance from the teacher on a Cartesian plane of the classroom	776
Row	Student's row in the classroom	776
Detection percentage	Percentage of the recorded time that the student was detected	668
Head travel	Accumulated changes (deltas) of the head horizontal rotations over time.	496
Head travel (norm.)	Head travel normalized over the measurements of the specific person in the class.	482
Number of still periods	Number of periods (of minimal duration of 5 seconds) during which the head movement can be considered still	668
Mean still period duration	Mean duration of the still period (as defined in the previous row)	618
Still time percentage	Percentage of time within the attention period during which the head was still.	668
Attention	Reported level of attention (1–10)	715
Attention labelled	Attention reports mapped to categories <i>low</i> , <i>medium</i> , <i>high</i>	715

Table 7.3 – Features used for the behaviour analysis of the individuals. First three features represent general information connected to spatial location and time of the class. Following six features were extracted from our observations of people's behaviour. Final two are the levels of attention which we will try to predict.

For the purposes of predicting student's attention level, we also included the spatial and time-based features. Location of the student was described with the Cartesian distance of the student from the teacher and the row number (two features which were identified as significant influences in Section 4.4.2.1). Even though time did not have a significant influence on attention we included the period of the class as a feature in our prediction attempts. Overview of all features is given in Table 7.3.

7.2.2 Modelling behaviour over time

In order to try to capture the direct synchronization between the teacher as the signal emitter and student as the receiver of information, we formed a hypothesis that the people with higher synchronization will also have a higher attention level. In order to test this, we modelled the interaction between the two measures (teacher's position and student's head orientation) with a number of measurements.

Using the maximal sampling rate, the head orientation and teacher's position data was sam-

pled 24 times a second (every frame). Given that we are sampling attention in intervals of approximately 10 minutes, we have to aggregate the 14400 ($10mins * 60sec * 24frames$) samples of students behaviour into a single value comparable with attention. Periods with less than 25% of maximal number of samples were rejected as invalid.

We already proposed our function for estimating gaze uncertainty from head pose in Section 6.5. The function models the expected limits of gaze direction based on the observed head pose. In order to test the correlation between the student's head motion and teacher's location in the classroom, we modelled the gaze behaviour with a number of functions. In the increasing order, each model adds additional assumptions about the gaze behaviour. All models are visualized in Figure 7.13.

Step function returns 1 if the teacher lies within the gaze limits of the student and otherwise 0. The function disregards the behaviour of eyes completely, and shows just whether the teacher is visible to the student based on the head orientation. The values collected over 10 minutes are aggregated by calculating the mean, which effectively represent the percentage of time the teacher spent in the view-field of the student.

T-H Correlation assumes that there is a linear relationship between the head orientation (centre of the view-field) and location of the teacher. We calculated the Pearson's r coefficient of correlation between the teacher's location and projection of student's head orientation on the front wall. A higher correlation coefficient indicates that the head movement follows the teacher's movement.

T-H Distance has a similar assumption, that the students with higher attention will more closely follow the movement of the professor. To measure this, at every time point we calculated the distance between the teacher's location and projection of student's head orientation on the front wall. Mean distance over the period of 10 minutes as used as the final value.

Mean pose prediction (MPP). We have shown in Section 6.4.1 that the usage of gaze within the limits is similar to a 2D Gaussian distribution positioned in the centre of the view-field. Here, we are using this assumption only for the horizontal plane and model probability of horizontal gaze direction with a normal distribution. The probability is centred on the projection of the centre of view-field and standard deviation is equal to half of the view-field width (view-field limits account for 95% of cumulative probability). The model shows how predictive is the student's gaze of the teacher's location. Probability values for teacher's locations over the whole period are averaged into a single value.

Normalized mean pose prediction (nMPP). Formulating the probability based on the view-field angle implicitly makes the probability scores for people sitting in the back lower (the distribution is "wider" and "flatter"). In order to cancel out this bias, we normalized the probability values with the peak-value of the probability when looking at the closer edge of the projection area. This put the output values in 0.0 - 1.0 range. Mean was used to aggregate values into a single measure.

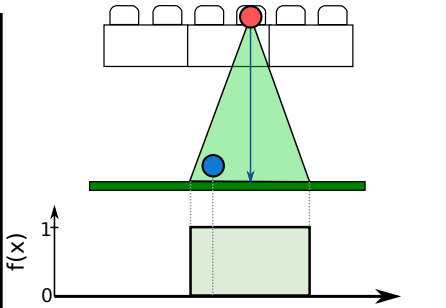
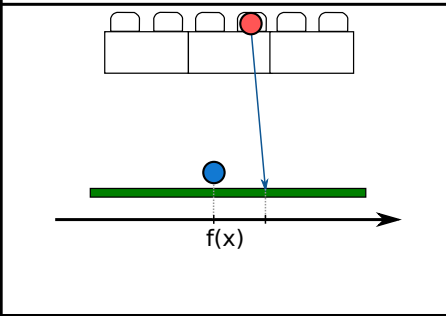
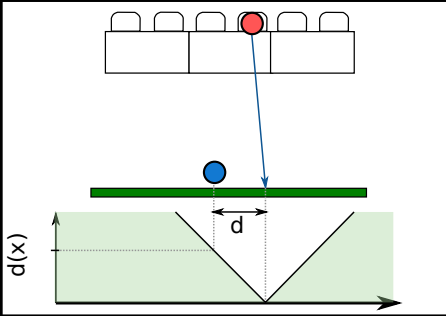
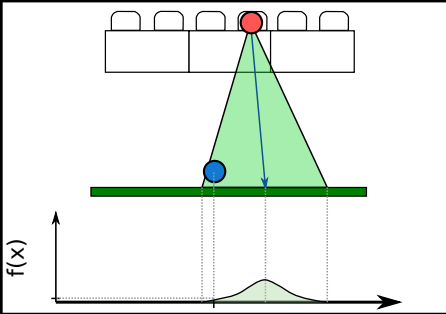
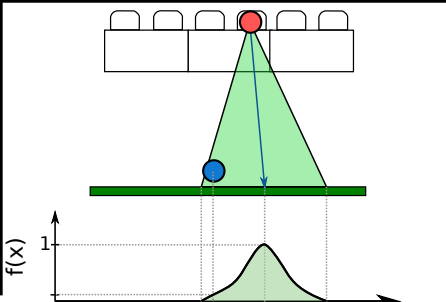
<p>Step function</p>		<p>Aggregated value: mean</p> <p>Practical meaning: Percentage of time the teacher spent in the student's view-field</p>
<p>T-H Correlation</p>		<p>Aggregated value: Pearson's R between pairs of teacher's position and student's gaze projection</p> <p>Practical meaning: The correlation level between two values.</p>
<p>T-H Distance</p>		<p>Aggregated value: Mean distance between teacher's position and student's gaze projection</p> <p>Practical meaning: How closely does the student's head orientation follow the teacher position.</p>
<p>Mean position prediction (MPP)</p>		<p>Aggregated value: Mean value of teacher's position probability in the student's gaze distribution.</p> <p>Practical meaning: Student's gaze predictability of teacher's location.</p>
<p>Normalized mean position prediction (nMPP)</p>		<p>Aggregated value: Mean value of normalized teacher's position probability in the student's gaze distribution.</p> <p>Practical meaning: Student's gaze predictability of teacher's location, with effects of distance neutralized.</p>

Figure 7.13 – Visualizations of the gaze modelling methods used. Red dot illustrates the student, with blue arrow line showing the centre of view-field and green triangle depicting the field of view. Blue dot represents the teacher's position in the front of the classroom. We illustrate measures used in each method and the output used below the sketch of the classroom.

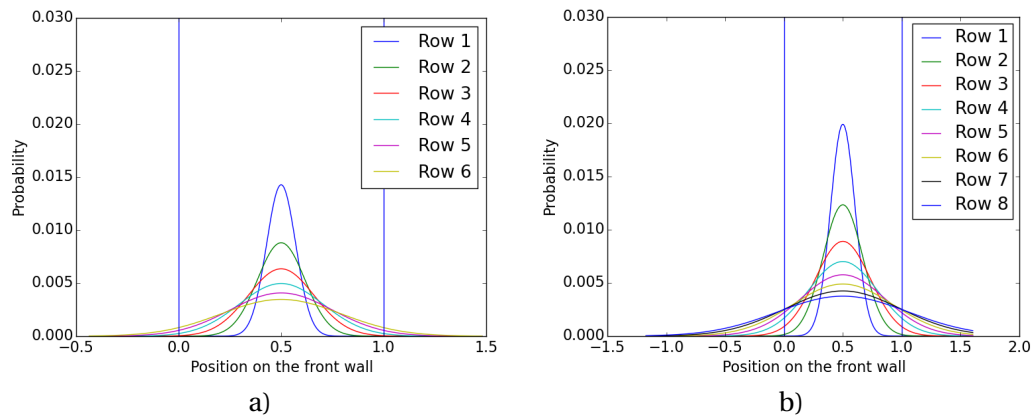


Figure 7.14 – Shape of Gaussian probabilities modelling gaze probability for the student located in the centre of “projection zone” ($x=0.5$) sitting in different rows. As a reference, vertical lines represent the edges of “projection zone” (0.0 - 1.0 span in the normalized positional coordinates). Models are given for **a)** Classroom A and **b)** Classroom B.

7.2.3 Influence of distance on field of view

Students in the back of the classroom have wider view-field than the students in the front. With this, they typically need to move their head less to maintain the visual contact with the teacher. We illustrated the effect of distance on the Gaussian that models the gaze probability within the gaze limits in Figure 7.14.

From our studies in head participation (Section 6.4.2.1), we found that in controlled settings head participation is present in all horizontal angles. This gives us theoretical basis to think that even though teacher’s motion will be observed in the back rows as a smaller angular displacement, the head movement should still be present.

We tried to model the predictive power of a person sitting at certain location in the classroom in a number of ways, trying to base them on:

- angular width of the projection zone and its relationship with the angular width of gaze limits;
- gaze probability function and its increased variance for the back rows.

In all cases, our function represented a slight modification of the “distance from the centre of projection zone” which was already included in our prediction models. Our other attempt to eliminate the influence of distance was already mentioned in the previous section, where we introduced the nMPP (normalized version of the mean pose prediction).

7.3 Research questions

Our research focused on the set of three main questions about the relationship between the teacher's actions and student's reactions.

1. Analysing students individually, is there a relationship between the individual's head-movement measurements and reported attention levels? If so, can we use the measurements to predict the attention level of the student?
2. We assume a relationship between teacher's position and student's head orientation. In order for this measurement to be predictive, we should find out how much did the teachers move in our recordings and what were the typical standing locations in the teacher's area?
3. Finally, what are the predictive power of our gaze models on student's attention and can we successfully use gaze limits introduced in the previous chapter for our classroom studies?

7.4 Results

7.4.1 Individual measurement analysis

7.4.1.1 Measurement performance

First significance tests showed the correlation between the attention level (on the scale 1-10) with the percent of time the person was detected (Pearson's $r(575) = 0.1158$, $p = 0.01$). This can be explained with the idea that engaged students will maintain more contact with the activities in the classroom. Apart from being more visible, student's head travel did not show significant difference on the overall scale. We expected this, as the measurement itself can be easily affected by noisy measurement.

After eliminating the individual differences with normalization of head travel, we found that positive changes in attention were reflected in increase in head travel (Pearson's $r(234) = 0.21$ $p = 0.0011$), shown in Figure 7.15. The observation, however, relies on the the change of attention within one lecture, and would not be able to correctly assess students who are constantly at a high level of attention.

Of the measures of stillness, only "percentage of time spent still" recorded a significant, but very weak correlation (Pearson's $r(614) = 0.09$, $p = 0.02$). After comparing it with the "percentage of time detected" we found a very high and significant correlation between the two measures ($r(666) = 0.91$, $p < .0001$). This does not allow for great significance of the second measure, because "percentage of time detected" is easier to calculate and with stronger correlation with attention. We kept all measures for further testing in our predictive models.

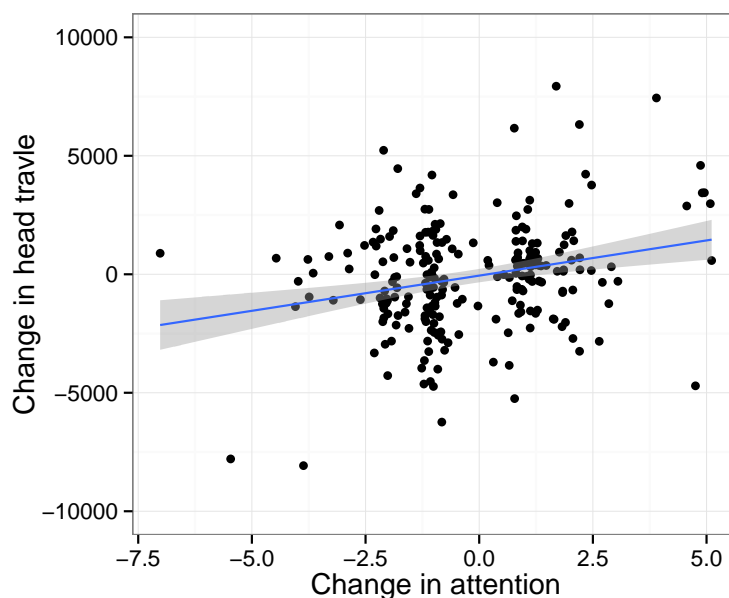


Figure 7.15 – Change in normalized head travel correlated to the change in attention. Red line represents the linear fit. Pearson's $r(204) = 0.21$ $p = 0.0011$. Gaussian noise added for the purpose of visualizing the samples without overlap after the linear fit.

7.4.1.2 Predictive model of attention

The supervised training was designed in order to explore the features further and evaluate their joint predicting accuracy. For this reason we did an exhaustive search of all feature combinations and SVM parameters to achieve the best prediction of the three categories of “labelled attention” – *low* (100 samples), *medium* (270 samples), *high* (246 samples).

The training procedure included a 64–16–20 split (64% of the data used for training, 16% for testing the parameters during the training, 20% for the final evaluation of the classifier) to find the best input combinations. Instead of the initial 80–20 split, we choose to split the 80% of data used for training again in the 80–20 ratio in order to avoid indirect over-fitting the classifier on the whole dataset (the last 20% were never used before the final evaluation).

We iterated over the parameters of the SVM (kernel type - *linear*, *polynomial*, *rbf*, and relevant parameters for each kernel), with gradual narrowing down of the parameter sampling step (step sizes were narrowed down in sequence 0.1, 0.01, 0.001). Four best scoring classifiers are given in Table 7.4.

Our concern was that the main informative source would rely on the *Detection percentage* or *Percentage still*, the two being highly correlated. This did happen in the earlier training attempts, but the features are not represented in the final set of classifiers (*Detection percentage* is used in the 10th best classifier). All of the best classifiers included a similar mix of features – head motion representatives, and some indications of distance and time of the class.

Kernel	Features	Score	Cohen's κ
RBF(c=1.31, g=0.0211)	Distance, Head travel norm., Num. still periods	61.86%	0.30
RBF(c=1.21, g=0.11)	Period, Row, Head travel norm., Mean duration still	61.72%	0.32
RBF(c=1.11, g=0.061)	Head travel norm., Mean duration still	60.42%	0.28
RBF(c=1.4, g=0.04)	Period, Distance, Row, Mean duration still	59.23%	0.30

Table 7.4 – Classifier scores for predicting “attention labelled”. Score given represent the prediction score on the 20% test sample. Parameters of the kernels are abbreviated as c - penalty for the error term; g - gamma.

Normalized head-travel measurements and Mean duration of still periods appears to be the most salient feature (both used in 3 of the 4 detectors).

With the best result of 61.86% correct estimations (Cohen's $\kappa = 0.30$) on an independent test set, our prediction scores, although not perfect, are high above chance. This confirms that we have constructed a set of features that is indicative of student's attention. We will demonstrate that we can further improve our results in the following sections, by using features based on teacher's motion.

7.4.2 Teacher's movement analysis

We used our tracking data to extract basic information about teacher's motion and usage of space. An raw output of the tracker over time is displayed in Figure 7.16. To illustrate better the fact that teachers do spent most of the time in the vicinity of the “projection zone” (limits are represented as the blue region in the figure) we calculated the percentage of time spent in each positional bin (explained in Figure 7.17a). We analysed but found only minor difference in the usage of space between the two teachers recorded.

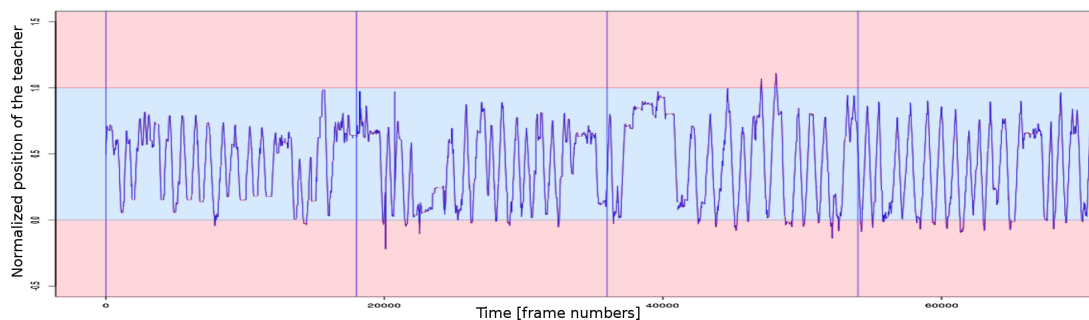


Figure 7.16 – Visualization of teacher's tracking. Horizontal axis represent time (vertical lines designate 10 minute intervals). Coloured regions are used to represent the “projection” (blue) and “outer” (red) zone of the teacher's area (as explained in Figure 7.12).

We did a simple classification of data points as “moving” or “standing”. A data point was

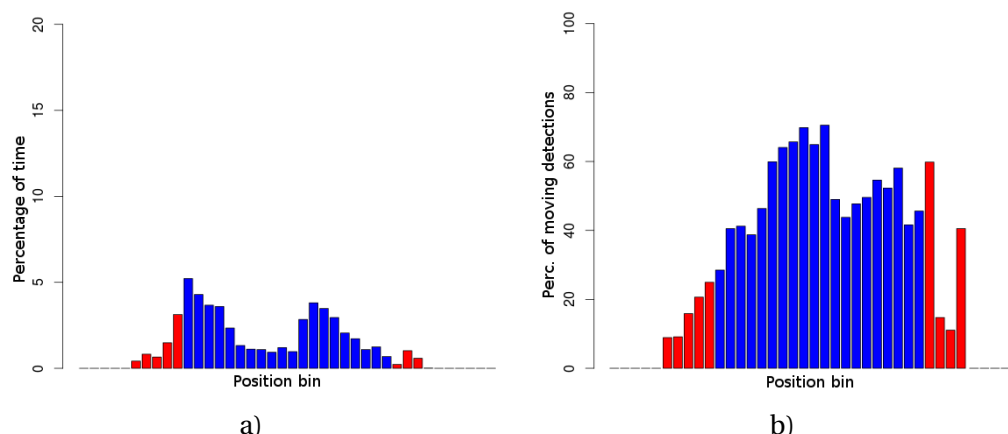


Figure 7.17 – Motion and standing positions of recorded teachers. **a)** Percentage of time spent in each of the positional bins in all recordings. **b)** Percentage of all detections classified as “moving” per bin.

classified as “moving” if the sliding window standard deviation was higher than 25% of the overall standard deviation (window size used was 5 measurements). This threshold gave us provisional values that in the recorded sample Teacher 1 spent 50.82% and Teacher 2 spent 23.27% of the time moving.

To illustrate which areas were more “transitional” and which were used more for standing, Figure 7.17b shows the percentage of observations classified as “moving” per positional bin. We can see that due to the usage of the projector, the centre of the projection zone was mostly transient and teachers tended to stand mostly towards the edges of the projection zone. The moving/standing classification was not used in further results.

The analysis of teacher’s movement shows us that teachers in our sample occupy at least two distinct locations for approximately same percentage of time (left and right edge of the projection zone) and that there is a reasonable amount of movement present in all recordings. Although a sample of two teachers does not allow general conclusions about the behaviour of teaching professionals, it gives ground for the usage of head-tracking measurements in our experiments.

7.4.3 Connection between student’s gaze and teacher’s position

7.4.3.1 Measurement’s performance

All used gaze models had significant predictive capabilities of student’s attention level. Table 7.5 shows the results of Spearman’s correlation between reported attentions and used gaze models, and relationships are visualized in Figure 7.18.

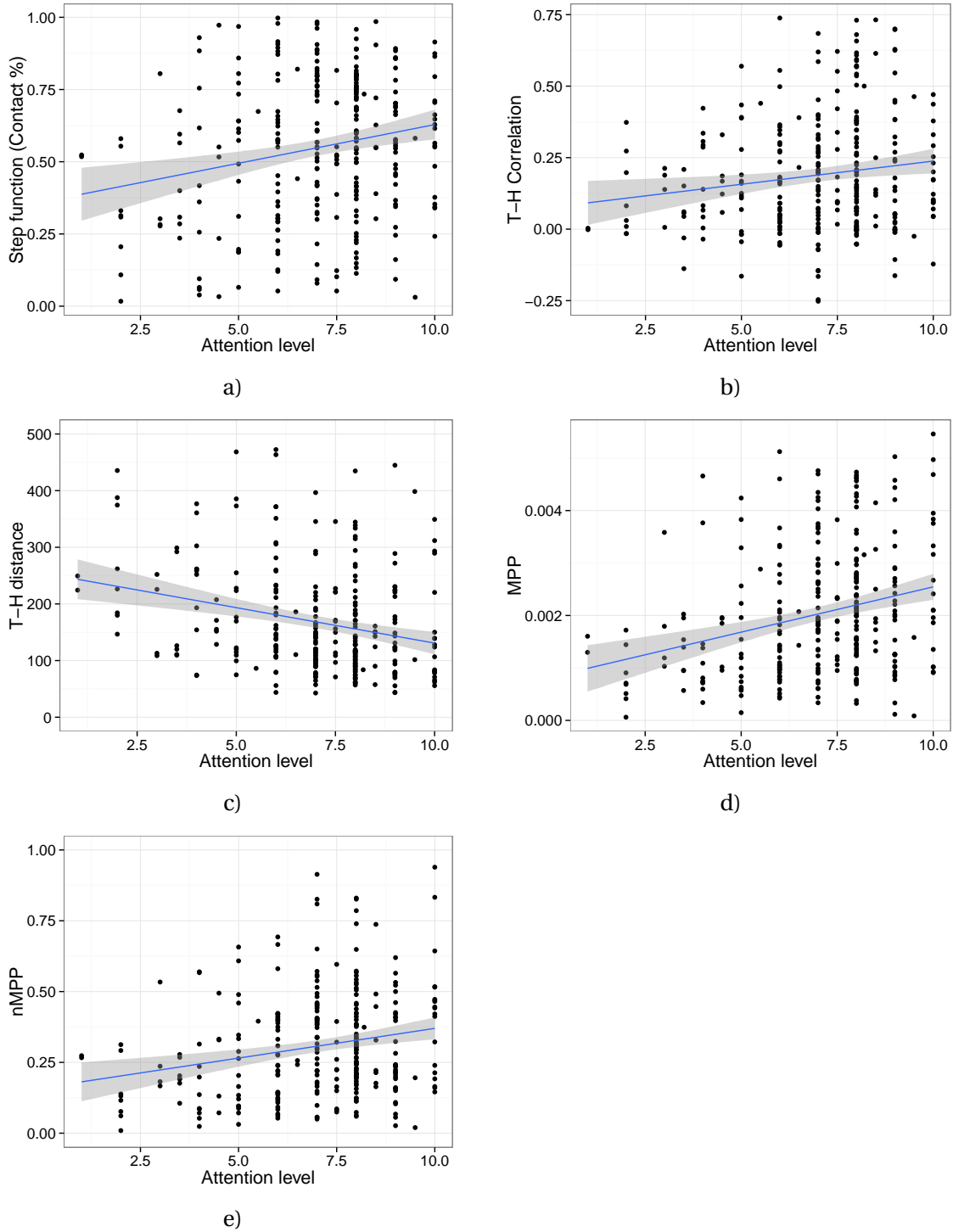


Figure 7.18 – Relationship between different models of gaze and reported attention of students. **a)** Step function, **b)** T-H correlation, **c)** T-H distance between gaze projection and teacher's position, **d)** Mean (teacher's) Position Predictability (MPP) - how much does the gaze predict the location of the teacher and **e)** normalized version of the same measurement (nMPP) - values for each student normalized to the range of 0.0-1.0.

Gaze model	Spearman's ρ	df	p-value
Step function	0.17	289	0.0045
T-H correlation	0.13	249	0.0422
T-H distance	-0.23	289	0.0001
MPP	0.27	289	<.0001
nMPP	0.18	289	0.0017

Table 7.5 – Result of Spearman's correlation between the temporal measures modelling contact between the gaze of the student and position of the teacher over periods of 10 minutes.

Using T-H correlation to model the relationship between the projection of student gaze on the front wall and teacher's position turned out to be the worst performing model. Even though the correlation is still significant we can conclude that it was a bad abstraction of the relationship between motion and gaze.

Step function was the second worst model, but still performed very well. It is interesting to notice that the measure favoured people with broader view-field (e.g. people in the back of the classroom), given that they had higher chance of overlapping with the teacher's position.

T-H distance between the projection of the student's gaze and teacher's position was the second model (apart from correlation) not to take into consideration the gaze-limits. The measurement has a negative relationship with attention, which was expected - the closer the student follows the teacher, the smaller the distance.

Mean position predictability (MPP) and its normalized variation (nMPP) both performed well, MPP having the strongest correlation with the attention among the models. We can conclude that measuring how well does the student's gaze at predicting teacher's location is a good measure to model the student's behaviour. At the same time, this model had the most complete set of assumptions about the student's gaze, and gives additional validity to our gaze usage model from Chapter 6.

It is interesting to see that the normalized MPP, which tried to neutralize the distance and by normalizing the functions gave advantage to the people in the back still performed reasonably well.

As our samples were collected at every frame, we explored whether it is possible to collect data at smaller sampling rates (i.e. bigger time steps) in order to lower the processing demands needed for data collection. We tested sampling the data in intervals of 1 second, 2 seconds, 5 seconds, 10 seconds and 30 seconds. Within the sampling interval we used mean values of defined measures to form a single measurement. Similarly, we used mean and median to approximate teacher's location within the time step (there was no significant difference in the results between the two aggregations). Values are given in Table 7.6.

We see that time steps initially have a beneficial effect on most functions. We speculate that this is caused by the additional smoothing of data which removed some of the noise. Because

Samp. step	Step func.	T-H correl.	T-H dist.	MPP	nMPP
1 sec	0.18 / 0.0004	0.13 / 0.0046	-0.27 / <.0001	0.30 / <.0001	0.22 / <.0001
2 sec	0.18 / 0.0004	0.18 / 0.0077	-0.26 / <.0001	0.30 / <.0001	0.23 / <.0001
5 sec	0.18 / 0.0003	0.11 / 0.11	-0.26 / <.0001	0.29 / <.0001	0.23 / <.0001
10 sec	0.16 / 0.0013	0.13 / 0.09	-0.25 / <.0001	0.28 / <.0001	0.22 / <.0001
30 sec	0.18 / 0.0004	0.07 / 0.49	-0.26 / <.0001	0.28 / <.0001	0.22 / <.0001

Table 7.6 – Result of Spearman's correlation (ρ value / p -value) over different sampling time-steps. Values within each sampling step was aggregated into a single measure by doing mean over the data collected within the time period.

we still keep the limit that a valid sample needs to have minimum 25% of values, bigger time-steps also lowered the number of data samples which caused the gradual weakening of correlations. In case of T-H correlation, the overall relationship with attention becomes insignificant at sampling steps above 5 seconds. Step function showed no changes because different sampling steps resulted in equivalent operations to the 1-frame data collection (doing mean-of-means), and the only difference displayed can be potentially explained by some samples being rejected in the 10sec sampling step.

Building on the idea that relationship of gaze with the teacher's position showed predictive capabilities, we tried including the projection area into our considerations. At every time point, the gaze was analysed as being in contact with either **i)** teacher, **ii)** projection area or **iii)** other directions. Even though looking in other directions was negatively correlated with attention (Spearman's $\rho(374) = -0.14$, $p = 0.0087$, sample rate 1 frame), teacher/projection contact measures did not show a conclusive correlation with attention. Our conclusion was that we encountered the same problem as Voit and Stiefelhagen (2010), and that in cases when teacher was standing in front of the projection area our modelling could not reliably differentiate between the two.

7.4.3.2 Attention prediction

Our next step was to try to use the new features in a predictive model for estimating student's attention level. We used the same training principle as previously described in Section 7.4.1.2 and again trained our SVM classifier to predict labelled student's attention (3 levels: low, medium and high). We used all of the features extracted about the individuals behaviour and our new features modelling the synchronization between the student and the teacher. Results are given in Table 7.7.

We see that the introduction of gaze metrics raises the prediction scores of attention around 10%. We are reporting four best-scoring predictors from the testing of all feature combinations with random assignment of samples to train and test-set. With the score of 70% of correct attention-level classifications and $\kappa = 0.47$, we can demonstrate that our features achieve significant results at predicting attention of students with standard ML approaches.

Kernel	Features	Score	Cohen's κ
RBF(c=1.4, g=0.4)	Distance, Row, Mean duration still, Step func.	71.74%	0.47
RBF(c=1.3, g=0.2)	Distance, Step func, T-H Correl	70.73%	0.46
RBF(c=1.4, g=0.3)	Distance, Row, Mean duration still, Step func.	69.56%	0.43
RBF(c=1.4, g=0.3)	Period, Distance, Row, Head travel norm., Perc still, MPP, Step func.	69.38%	0.20

Table 7.7 – Classifier scores for predicting “attention labelled”. Score given represent the prediction score on the held-out 20% test sample. Parameters of the kernels are abbreviated as c - penalty for the error term; g - gamma.

7.5 Conclusion

In this chapter we presented in detail our methods for extracting data about human gaze and head behaviour. We also discussed the possible problematic scenarios for data extraction from a large group of closely positioned individuals in classrooms.

A number of measurements connected to the head-movement of individuals was defined. Even though the measurements individually demonstrated modest effects on predicting attention level, we were able to combine them into a well-performing predictive model. Finally, using our hypothesis that there is a connection between student's attention and behaviour, we have defined a number of models of relationship between the student's gaze and teacher's motion. Our models show that students with higher correlation to teacher's actions can be identified as being more attentive. Best predicting model of this behaviour (MPP, $\rho(289) = 0.27, p = < .0001$) was based on our head-to-gaze modelling presented in Chapter 6. By showing that even the broadest assumptions (step function, with $\rho = 0.18$) showed predictive power, we also hinted that the proposed methods can work with limited performance even without the fine-grained measurements such as the exact gaze direction.

By using all of the head and gaze features, combined with general spatial information, we have achieved scores of around 70% correct classifications (and $\kappa \approx 0.45$) on a 3-point scale of attention.

We do not claim that we exhausted the list of available cues about human interaction with this study, or that our methods set the new state-of-the-art, but we see this study as the proof-of-concept for a novel view of classroom activities. Unobtrusive behaviour analysis is demonstrated as a valid source of attention prediction, which can easily be scaled to large number of students.

Features used here fit in our general theory of direct synchronization between the source of information (teacher) and the receiver (students). From the estimation of individual attentions, we can formulate a single measure of classroom entropy based on collected data about the individual students. The reason why this was not directly addressed in our study is because the mean attention of the classroom showed little variations, and measure of individual's

Chapter 7. Gaze in the classroom

attention was considered more informative (variance of mean attention on the scale from 1-10 was 0.26 in our sample).

Our technical conclusions can be also be mirrored on the pedagogical side. With distinctive actions, such as distributed spatial presence and frequent interactions, teachers can easily conduct informal queries of attention. Even without a formal framework, this can serve as a base for developing into a “reflective practitioner”.

8 Conclusions

DURING the lecture, teachers need to balance their attention between the lecture material, their presentation and audience's reception. Even with perfect execution, the lecture fails if the students are not attentive. Our work explored whether we can unobtrusively collect information about the student's attention, by analysing non-verbal behaviour with computer vision algorithms. We explored in detail features of student's body motion and head movement as an approximation of the gaze direction. We will now give a short overview of the contributions and empirical results presented in previous chapters. We will then discuss the known limitations of our studies, and highlight potential directions for future research.

8.1 Summary

We aim to estimate student's level of attention during the lecture, with the idea that there are behavioural cues that experienced lecturers have been using to adjust their presentation to the audience's mood. The goal of this thesis was to try to capture such behaviour automatically by:

- formulating a theoretical framework that would explain the non-verbal behaviour,
- defining a set of unobtrusive metrics which would capture it,
- and develop a system for experimental validation of our ideas on a collected data-set.

Main principles in our approach were, under the guidance of social sensing and unobtrusive measurements:

- that the methods developed should not disrupt existing teaching or learning practices,
- that data collected should reflect realistic situations as much as possible,
- that social signals sent unconsciously will probably be more honest than the ones purposefully displayed.

8.1.1 Results

We conducted a series of recordings with four student populations collecting over 340 questionnaires, 30 hours of video material and numerous other data sources (interviews, eye-trackers, etc). In our analysis we found that:

- Attention of students is spatially influenced. Our results from the questionnaires have shown that the back rows reported lower levels of attention than the front ones. Our motion study has shown that movement synchronization is more likely to happen between neighbours, which leads to a conclusion that the immediate neighbours will influence each other even in non-collaborative activities. Finally, studying students' seating habits over the semester has shown us that students will not adjust their seating habits to their current mood, but are more likely to act in accordance with their habits.
- Based on the idea of “indirect synchronization” between audience members, we were able to differentiate between the higher- and lower-attentive audience members. The measurement called “motion lag” was defined with the idea that the less attentive students would be slower to react on the same stimulus compared to their visible neighbours. The main properties of the measurement are that it is independent from the presented content and individual's personality traits.
- Our controlled study of the relationship between gaze and head movements further confirmed previous models based on the linear relationship between the two. Even though head orientation was shown to be of modest participation, it was present over all tested angles regardless of their intensity. In addition, our real-world recordings showed that the gaze patterns are highly concentrated around the centre of the view-field, which leads us to believe that over long periods of time head orientation is indicative of the direction of gaze.
- We formulated and explored a number of measurements connected with the head-motion of students during the lecture. De-contextualized measurements, measuring only the person's activity, showed correlation with attention, but the real benefits of the measures were observed when we compared them against the recorded motion of the teacher. Our experiments demonstrated that higher-attentive students followed the actions of the teacher more closely with their estimated gaze. The observed effect is considered as an example of “direct synchronization” between the source (teacher) and the receiver (student) of informations. Finally, we demonstrated the predictive power of the proposed measures by training a model of student's attention based on standard machine-learning algorithms.

Our final conclusion is that student's behaviour in class is rich with cues about their attention. While some of the indicators are visible to human perception (head movement), others are beyond our causal observations (synchronization of movement). Proposed formalization and

automatic processing of such indicators allow us to easily scale with the size of the audience, which is shown to be a difficult task for human perception.

8.1.2 Contributions

The main contribution of this work is the non-exhaustive set of defined measurements of visually observable human behaviour for assessing audience attention in the context of a lecture. With the exploratory nature of this research, we do not claim that the list is absolute, or that all directions are fully exhausted. The main advantages of the proposed set, in accordance with unobtrusiveness principle are that these metrics are:

- independent of the lecture's content,
- either identity-independent or connected to the person only to the extent of a single lecture,
- require no changes in the teachers' approach to their students,
- require no manual intervention for their execution, neither from the teacher nor the students.

We consider these properties important in order for the approach to be implementable without privacy-invasion and without additional operational overhead, thus reaching real classrooms one day.

Our second contribution is the unified theory about the underlying principles behind these measurements. This theory, briefly presented in Chapter 3, allows for expansions on the set of proposed metrics, by formulating what we believe are some of the principles behind the complex topic of human communications.

Our controlled experiment on the connection between gaze and head orientation is the most formal attempt at evaluating the relationship, to the best of our knowledge. The final formulation of the gaze uncertainty model unifies a number of observations about real-world gaze usage in a complete, reusable model.

As it is customary with multi-disciplinary research, a number of minor contributions are connected to the techniques used. Our detailed description of the processing pipe-lines serves both as a list of observed pit-falls in feature extraction and attempts at solving them – such as our formulation of *motion tracks* and observations about the properties of head motion and detection assignment. The collected dataset of video recordings and attention data is also unique to our knowledge, and will hopefully be reused in future research.

8.2 Limitations

A number of assumptions are connected with our findings. Potentially the most limiting is the format of the lecture being analysed. Teaching is a versatile activity, and the number of pedagogical techniques proposed is growing every year. Presented findings will not be equally valid across all of them. Our focus was on the scenario which we consider the most frequent. This relates to both the presentation style (lecturing with slides) and to the geometry of the student arrangement.

Secondly, the focus of our exploration was on the relationship between the teacher and students, based on our hypothesis that the teacher will be the dominant source of information during a lecture. As we previously stated, a number of other sources are competing for students' attention, and in order to completely capture the learning experience we should analyse all of them in a holistic way – e.g. including information on what is on the student's desk and what is being shown on the slides. The opposite direction is also valid, and more commonly pursued – capturing the learning in a controlled environment would raise the internal validity of our findings, at the danger of mis-representing the experience (lower ecological validity).

Also connected to ecological validity, we base our results on the students' reports of attention levels. Even though we tried to make sure that the students were free to openly express their opinions, the collected data remains implicitly subjective. Because other teaching approaches might be more effective, we also refrained from making the jump from attention to learning gains, although a strong connection is indicated by previous literature. We tried to adjust our data-collection for minimal influence on the students' perception of the lecture. Although necessary steps were taken for familiarizing the students with the experiment and collected opinions were generally positive, further improvements can be made with less conspicuous recording setups. Additional validation of metrics can also be done by including a number of teachers from a range of subjects, and different levels of professional experience.

Further analysis of the teacher's actions are desirable for capturing the full picture about the communication signals. Information about the gaze of the teacher and its influence on student's attention was planned, but not implemented because of technical problems of extracting meaningful data. Of course, teacher's tools are not limited to the gaze. Gesture analysis in general could also be beneficial to extract further information about teaching style, and connect that with the students' response.

Range of audience sizes covered by our measures should be further tested. We expect that the behaviour of the audience changes with the perception of the group, and our samples were based on typical class sizes (covering roughly the 20-50 people range). As previous research has shown, small classes (less than 15 students) have different, more favourable social dynamics between audience and teachers. Because of that, we do not expect that our observations would be effective (or needed) in small groups. We have not tested our approach in university auditoriums (100+ students) because of the difficulty we had finding a teacher and audience of that size willing to participate. Bigger audiences represent an excellent challenge for teacher's

presentations skills, and would possibly emphasize benefits of our intervention.

Finally, additional improvements can be done to our feature-extraction techniques. Because of time restrictions, some of the validation steps were simplified, and should be re-done with more reliable ground-truths and larger sample collections. Given the rapid pace of Computer Vision development, we are confident that better-performing face-detection/pose-estimation algorithms will be available by time this thesis is published.

8.3 Future work

Our work represents one of the attempts of bringing learning analytics to the non-digital world. While keeping all the benefits of the digital-domain approaches (easy data collection, big data samples, clear data meaning) and aiming at the dominant teaching scenario, it shows great potential, but not without challenges.

A number of additional information sources from the classroom can be collected. At the early stages of our research, we took into consideration the data presented on the projector by annotating the changes and properties of the presented information, before our focus settled on the human relationships. Similarly, attempts to collect more objective measures of attention were considered e.g. by using consumer-level EEG monitors. Even though this would potentially increase the intrusiveness of the approach, we might have been able to study the changes of attention at a much higher resolution than the 10-minute sampling used for our research.

As shown by our interviews, the voice of the teacher has a strong influence over the students, and is considered one of the main characteristics of the presenter. We have no doubt that classroom dynamics modelling can be enhanced by including other signal modalities beyond vision. Similarly, direct interaction between the teacher and students (e.g. question-answer episodes), and analysing student's active participation in the learning process would open a whole new set of social cues. We expect that this dynamic would be of higher complexity, and aside from the problems connected with the extraction of valid data, social-network analysis and identity-based information would add depth to this line of research.

After establishing the metrics, the natural next step is to use them and see how they influence teaching. Different presentation of our observations can encourage both reflection-in-action and reflection-on-action. Some of interventions to be considered are: the teacher receiving a notification during the lecture if the attention of the audience is too low; teacher sitting down to compare his/hers view of the lecture with the collected metrics after class; or giving a time-line of class attention, indicating which parts of the lecture were captured with low attention and worth repeating next time.

In a broader aspect, proposed principles may also be valid for other audiences. It would be interesting to explore how our metrics behave in conferences, theatres or cinema. Currently

Chapter 8. Conclusions

used technologies have a jump between collecting broad demographics for assessing television channel ratings, to using facial analysis of an individual in a controlled viewing. The fact that information can be assessed on very large and imprecise samples or on a very small sample leaves a lot of room for our class of metrics.

As with other forms of human interaction, deciding future directions of research is not limited by the number of information we can extract from the activity but when do the collection costs out-weight the benefits. We consider our current explorations to be well balanced in this aspect. Human contact should, after all, remain human.

Bibliography

- Raymond S Adams. Location as a feature of instructional interaction. *Merrill-Palmer Quarterly of Behavior and Development*, 15(4):309–321, 1969.
- Karl Albrecht. *Social Intelligence: The new science of success*. John Wiley & Sons, 2006.
- Irwin Altman and Evelyn E Lett. The ecology of interpersonal relationships: A classification system and conceptual model. *Social and psychological factors in stress*, pages 177–201, 1970.
- Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.
- Nalini Ambady, Frank J Bernieri, and Jennifer A Richeson. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in experimental social psychology*, 32:201–271, 2000.
- K. Anders Ericsson. Deliberate practice and acquisition of expert performance: a general overview. *Academic Emergency Medicine*, 15(11):988–994, 2008.
- John Robert Anderson, C Franklin Boyle, Robert Farrell, and Brian J Reiser. *Cognitive principles in the design of computer tutors*. Department of Psychology, Carnegie-Mellon University, 1984.
- Oya Aran and Daniel Gatica-Perez. Fusing audio-visual nonverbal cues to detect dominant people in group conversations. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3687–3690. IEEE, 2010.
- Michael Argyle. *Social interaction*, volume 103. Transaction Publishers, 1969.
- Michael Argyle. *Bodily communication*. Routledge, 2013.
- Ivon Arroyo, David G Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner, and Robert Christopherson. Emotion sensors go to school. In *AIED*, volume 200, pages 17–24, 2009.
- Stylianios Asteriadis, Kostas Karpouzis, and Stefanos Kollias. Visual focus of attention in non-calibrated environments using gaze estimation. *International journal of computer vision*, 107(3):293–316, 2014.

Bibliography

- Sileye O Ba and J-M Odobez. Recognizing visual focus of attention from head pose in natural meetings. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(1): 16–33, 2009.
- Sileye O Ba and Jean-Marc Odobez. A probabilistic framework for joint head tracking and pose estimation. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 264–267. IEEE, 2004.
- Sileye O Ba and Jean-Marc Odobez. A study on visual focus of attention recognition from head pose in a meeting room. In *Machine Learning for Multimodal Interaction*, pages 75–87. Springer, 2006.
- Sileye O Ba and Jean-Marc Odobez. Probabilistic head pose tracking evaluation in single and multiple camera setups. In *Multimodal Technologies for Perception of Humans*, pages 276–286. Springer, 2008a.
- Sileye O Ba and Jean-Marc Odobez. Visual focus of attention estimation from head pose posterior probability distributions. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 53–56. IEEE, 2008b.
- Sileye O Ba, Hayley Hung, and Jean-Marc Odobez. Visual activity context for focus of attention estimation in dynamic meetings. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1424–1427. IEEE, 2009.
- Khaled Bachour. *Augmenting face-to-face collaboration with low-resolution semi-ambient feedback*. PhD thesis, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, 2010.
- John L Barron, David J Fleet, and Steven S Beauchemin. Performance of optical flow techniques. *International journal of computer vision*, 12(1):43–77, 1994.
- Marian Stewart Bartlett, Gwen C Littlewort, Mark G Frank, and Kang Lee. Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology*, 24(7):738–743, 2014.
- Daniel J Bernstein. Peer review and evaluation of the intellectual work of teaching. *Change: The Magazine of Higher Learning*, 40(2):48–51, 2008.
- Sanjay Bilakhia, Stavros Petridis, and Maja Pantic. Audiovisual detection of behavioural mimicry. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 123–128. IEEE, 2013.
- Elina Birmingham, Walter F Bischof, and Alan Kingstone. Social attention and real-world scenes: The roles of action, competition and social content. *The Quarterly Journal of Experimental Psychology*, 61(7):986–998, 2008.
- Quentin Bonnard, Séverin Lemaignan, Guillaume Zufferey, Andrea Mazzei, Sébastien Cuendet, Nan Li, Ayberk Özgür, and Pierre Dillenbourg. Chilitags 2: Robust fiducial markers for augmented reality and robotics., 2013. URL <http://chili.epfl.ch/software>.

- Jean-Yves Bouguet. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 2001.
- Konstantinos Bousmalis, L Morency, and Maja Pantic. Modeling hidden dynamics of multi-modal cues for spontaneous agreement and disagreement recognition. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 746–752. IEEE, 2011.
- Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and vision computing*, 31(2):203–221, 2013.
- G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
- George Breed and Victoria Colaiuta. Looking, blinking, and sitting nonverbal dynamics in the classroom. *Journal of Communication*, 24(2):75–81, 1974.
- Hilton Bristow. Parts based detector in c++. <https://github.com/wg-perception/PartsBasedDetector>, 2012.
- Lawrence J Brunner. Smiles can be back channels. *Journal of Personality and Social Psychology*, 37(5):728, 1979.
- Jane E Caldwell. Clickers in the large classroom: current research and best-practice tips. *CBE-Life Sciences Education*, 6(1):9–20, 2007.
- Paul Cameron and Dorothy Giuntoli. Consciousness sampling in the college classroom or is anybody listening? *Intellect*, 101(2343):63–4, 1972.
- Nick Campbell. Multimodal processing of discourse information; the effect of synchrony. In *ISUC*, pages 12–15, 2008.
- Nick Campbell. *An audio-visual approach to measuring discourse synchrony in multimodal conversation data*. September, 2009.
- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *Machine learning for multimodal interaction*, pages 28–39. Springer, 2006.
- Kalina Christoff, Alan M Gordon, Jonathan Smallwood, Rachelle Smith, and Jonathan W Schooler. Experience sampling during fmri reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences*, 106(21): 8719–8724, 2009.
- Marvin M Chun and Jeremy M Wolfe. Chapter nine visual attention. *Blackwell Handbook of Sensation and Perception*, pages 272–311, 2001.

Bibliography

- Herbert H. Clark and Susan E. Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149, 1991.
- Doug Clow. Moocs and the funnel of participation. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 185–189. ACM, 2013.
- Robert Coe, Cesare Aloisi, Steve Higgins, and Lee Elliot Major. What makes great teaching? review of the underpinning research. 2014.
- Davide Conigliaro, Francesco Setti, Chiara Bassetti, Roberta Ferrario, and Marco Cristani. Attento: Attention observed for automated spectator crowd analysis. In *Human Behavior Understanding*, pages 102–111. Springer, 2013a.
- Davide Conigliaro, Francesco Setti, Chiara Bassetti, Roberta Ferrario, and Marco Cristani. Viewing the viewers: A novel challenge for automated crowd analysis. In *New Trends in Image Analysis and Processing-ICIAP 2013*, pages 517–526. Springer, 2013b.
- Davide Conigliaro, Francesco Setti, Chiara Bassetti, Roberta Ferrario, Marco Cristani, Paolo Rota, Nicola Conci, and Nicu Sebe. Observing attention. In *FISU Winter Universiade Conference*, 2013c.
- Mark Cook. Gaze and mutual gaze in social encounters: How long—and when—we look others" in the eye" is one of the main signals in nonverbal communication. *American Scientist*, pages 328–333, 1977.
- Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1): 38–59, 1995.
- Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.
- Roisin P Corcoran and Roland Tormey. *Developing Emotionally Competent Teachers; Emotional Intelligence and Pre-Service Teacher Education*. Peter Lang, 2012.
- M. Cristani, V. Murino, and Alessandro Vinciarelli. Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 51–58. IEEE, 2010. ISBN 1424470293.
- Mihaly Csikszentmihalyi. *Flow*. Springer, 2014.
- Jared R Curhan and Alex Pentland. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3): 802, 2007.
- Flavio Soares Correa Da Silva and Jaume Agusti-Cullell. *Information flow and knowledge sharing*, volume 2. Elsevier, 2008.

- Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.
- John A Daly and Amy Suite. Classroom seating choice and teacher perceptions of students. *The Journal of Experimental Educational*, pages 64–69, 1981.
- John Daniel. Making sense of moocs: Musings in a maze of myth, paradox and possibility. *Journal of Interactive Media in Education*, 3, 2012.
- Maria Danninger, Roel Vertegaal, Daniel P Siewiorek, and Aadil Mamuji. Using social geometry to manage interruptions and co-worker attention in office environments. In *Proceedings of Graphics Interface 2005*, pages 211–218. Canadian Human-Computer Communications Society, 2005.
- J Daum. Proxemics in the classroom: Speaker-subject distance and educational performance. In *annual meeting of Southeastern Psychological Association*, 1972.
- Barbara Gross Davis. *Tools for teaching*. Wiley.com, 2009.
- Emilie Delaherche and Mohamed Chetouani. Multimodal coordination: exploring relevant features and measures. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 47–52. ACM, 2010.
- Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *Affective Computing, IEEE Transactions on*, 3(3):349–365, 2012.
- Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516. ACM, 2013.
- Pierre Dillenbourg. Over-scripting cscl: The risks of blending collaborative learning with instructional design. *Three worlds of CSCL. Can we support CSCL?*, pages 61–91, 2002.
- Pierre Dillenbourg and Patrick Jermann. Technology for classroom orchestration. *New Science of Learning*, pages 525–552, 2010.
- Pierre Dillenbourg, Guillaume Zufferey, Hamed Alavi, Patrick Jermann, Son Do-Lenhand, Quentin Bonnard, Sébastien Cuendet, and Frédéric Kaplan. Classroom orchestration: The third circle of usability. In *International Conference on Computer Supported Collaborative Learning Proceedings*, pages 510–517. 9th International Conference on Computer Supported Collaborative Learning, 2011.
- Ahmet Cengizhan Dirican. A marker detection method using hysteresis thresholding for human posture tracking: a head tracking system. In *Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments*, page 36. ACM, 2014.

Bibliography

- Ahmet Cengizhan Dirican and Mehmet Göktürk. Involuntary postural responses of users as input to attentive computing systems: An investigation on head movements. *Computers in Human Behavior*, 28(5):1634–1647, 2012.
- Cosimo Distanto, Sebastiano Battiato, and Andrea Cavallaro. *Video Analytics for Audience Measurement*. Springer, 2014.
- Sidney D’mello and Arthur Graesser. Mind and body: Dialogue and posture for affect detection in learning environments. *Frontiers in Artificial Intelligence and Applications*, 158:161, 2007.
- Sidney D’Mello, Patrick Chipman, and Art Graesser. Posture as a predictor of learner’s affective engagement. In *Proceedings of the 29th annual cognitive science society*, volume 1, pages 905–910. Cognitive Science Society, Austin, TX, 2007.
- Vanessa Echeverría, Allan Avendaño, Katherine Chiluita, Aníbal Vásquez, and Xavier Ochoa. Presentation skills estimation based on video and kinect data analysis. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 53–60. ACM, 2014.
- Paul Ekman and Wallace V Friesen. Facial action coding system. 1977.
- Rana El Kaliouby and Peter Robinson. Mind reading machines: Automated inference of cognitive mental states from video. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 682–688. IEEE, 2004.
- Rana El Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*, pages 181–200. Springer, 2005.
- NJ Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, 2000.
- Edmund T Emmer and Laura M Stough. Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist*, 36(2):103–112, 2001.
- Gabriele Fanelli, Juergen Gall, and Luc Van Gool. Real time 3d head pose estimation: Recent achievements and future challenges. In *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on*, pages 1–4. IEEE, 2012.
- Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.
- Elizabeth Fennema, Penelope L Peterson, Thomas P Carpenter, and Cheryl A Lubinski. Teachers’ attributions and beliefs about girls, boys, and mathematics. *Educational Studies in Mathematics*, 21(1):55–69, 1990.

- Jeremy D Finn, Gina M Pannozzo, and Charles M Achilles. The “why’s” of class size: Student behavior in small classes. *Review of Educational Research*, 73(3):321–368, 2003.
- Martin A Fischler and Robert A Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on computers*, (1):67–92, 1973.
- Corporation Fitbit. On-line resources. <http://www.fitbit.com>, 2007.
- David A Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. Pearson, 2003.
- Donelson R Forsyth. *Group dynamics*. CengageBrain. com, 2009.
- Edward G Freedman and David L Sparks. Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. *Journal of Neurophysiology*, 77(5):2328–2348, 1997.
- Kenneth A Funes Mora, Laurent Nguyen, Daniel Gatica-Perez, and Jean-Marc Odobez. A semi-automated system for accurate gaze coding in natural dyadic interactions. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 87–90. ACM, 2013.
- Hua Gao, Anil Yuce, and Jean-Philippe Thiran. Detecting emotional stress from facial expressions for driving safety. In *International Conference on Image Processing (ICIP) 2014*, number EPFL-CONF-200407, 2014.
- Daniel Gatica-Perez. Analyzing group interactions in conversations: a review. In *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, pages 41–46. IEEE, 2006.
- Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775–1787, 2009.
- Daniel Gatica-Perez, Iain A McCowan, Dong Zhang, and Samy Bengio. Detecting group interest-level in meetings. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, number EPFL-CONF-83257, 2005.
- Titus Geerligs. Students’ thoughts during problem-based small-group discussions. *Instructional science*, 22(4):269–278, 1994.
- Tobias Gehrig and Hazım Kemal Ekenel. Why is facial expression analysis in the wild challenging? In *Proceedings of the 2013 on Emotion recognition in the wild challenge and workshop*, pages 9–16. ACM, 2013.
- Susan Goldin-Meadow, Debra Wein, and Cecilia Chang. Assessing knowledge through gesture: Using children’s hands to read their minds. *Cognition and Instruction*, 9(3):201–219, 1992.
- Nicolas Gourier, Daniela Hall, and James L Crowley. Estimating face orientation from robust detection of salient facial structures. In *FG Net Workshop on Visual Observation of Deictic Gestures*, pages 1–9. FGnet (IST-2000-26434) Cambridge, UK, 2004.

Bibliography

- Jonathan Gratch and Stacy Marsella. *Social emotions in nature and artifact*. Oxford University Press, 2013.
- Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.
- D Guitton and M Volle. Gaze control in humans: eye-head coordination during orienting movements to targets within and beyond the oculomotor range. *Journal of neurophysiology*, 58(3):427–459, 1987.
- Edward Twitchell Hall. *The hidden dimension*. Anchor Books New York, 1969.
- Andy Hargreaves. Mixed emotions: Teachers’ perceptions of their interactions with students. *Teaching and teacher education*, 16(8):811–826, 2000.
- John Hattie. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge, 2008.
- Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005.
- Corporation HeyDay. On-line resources. <http://hey.co>, 2013.
- Dirk Heylen. Challenges ahead: head movements and other social acts during conversations. 2005.
- Francis Heylighen. Complexity and information overload in society: why increasing efficiency leads to decreasing control. *Technological Forecasting and Social Change*, pages 1–19, 2004.
- B. Holmberg. *The Evolution, Principles and Practices of Distance Education*. ASF series. Bibliotheks- und Informationssystem der Univ., 2005. ISBN 9783814209333. URL <https://books.google.ch/books?id=YTtdNQAAAJ>.
- Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.
- Adrian Holzer, Sten Govaerts, Andrii Vozniuk, Bruno Kocher, and Denis Gillet. Speakup in the classroom: anonymous temporary social media for better interactions. In *CHI’14 Extended Abstracts on Human Factors in Computing Systems*, pages 1171–1176. ACM, 2014.
- Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. Models of attention in computing and communication: from principles to applications. *Communications of the ACM*, 46(3):52–59, 2003.
- Jay R Howard and Amanda L Henney. Student participation and instructor gender in the mixed-age college classroom. *Journal of Higher Education*, pages 384–405, 1998.

- Hayley Hung and Daniel Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *Multimedia, IEEE Transactions on*, 12(6):563–575, 2010.
- Allison M Jacobs, Benjamin Fransen, J Malcolm McCurry, Frederick WP Heckel, Alan R Wagner, and J Gregory Trafton. A preliminary system for recognizing boredom. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 299–300. ACM, 2009.
- JD Jecker, N Maccoby, and HS Breitrose. Improving accuracy in interpreting non-verbal cues of comprehension. *Psychology in the Schools*, 2(3):239–244, 1965.
- Patrick Jermann and Marc-Antoine Nüssli. Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1125–1134. ACM, 2012.
- Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.
- Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1409–1422, 2012.
- Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.
- Maria Karam. *PhD Thesis: A framework for research and design of gesture-based human-computer interactions*. PhD thesis, University of Southampton, 2006.
- Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.
- Eric I Knudsen. Fundamental components of attention. *Annu. Rev. Neurosci.*, 30:57–78, 2007.
- Martin Koestinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.
- Mele Koneya. Location and interaction in row-and-column seating arrangements. *Environment and Behavior*, 8(2):265–282, 1976.
- Vitomir Kovanović, Dragan Gašević, Shane Dawson, Srećko Joksimović, Ryan S Baker, and Marek Hatala. Penetrating the black box of time-on-task estimation. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 184–193. ACM, 2015.
- Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Bibliography

- Martin Lades, Jan C Vorbrüggen, Joachim Buhmann, Jörg Lange, Christoph V d Malsburg, Rolf Wurtz, and Wolfgang Konen. Distortion invariant object recognition in the dynamic link architecture. *Computers, IEEE Transactions on*, 42(3):300–311, 1993.
- Stephen RH Langton. The mutual influence of gaze and head orientation in the analysis of social attention direction. *The Quarterly Journal of Experimental Psychology: Section A*, 53(3):825–845, 2000.
- Stephen RH Langton, Roger J Watt, and Vicki Bruce. Do the eyes have it? cues to the direction of social attention. *Trends in cognitive sciences*, 4(2):50–59, 2000.
- Sophie I Lindquist and John P McLean. Daydreaming and its correlates in an educational environment. *Learning and Individual Differences*, 21(2):158–167, 2011.
- Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981.
- Bruce David Lucas. Generalized image matching by the method of differences. 1985.
- Gonzalo Luzardo, Bruno Guamán, Katherine Chiluiza, Jaime Castells, and Xavier Ochoa. Estimation of presentations skills based on slides and audio features. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 37–44. ACM, 2014.
- Alvaro Marcos-Ramiro, Daniel Pizarro-Perez, Marta Marron-Romera, Laurent Nguyen, and Daniel Gatica-Perez. Body communicative cue extraction for conversational analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.
- Helen M Marks. Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American educational research journal*, 37(1):153–184, 2000.
- Malia F Mason, Michael I Norton, John D Van Horn, Daniel M Wegner, Scott T Grafton, and C Neil Macrae. Wandering minds: the default network and stimulus-independent thought. *Science*, 315(5810):393–395, 2007.
- Marianne Schmid Mast, Daniel Gatica-Perez, Denise Frauendorfer, Laurent Nguyen, and Tanzeem Choudhury. Social sensing for psychology automated interpersonal behavior assessment. *Current Directions in Psychological Science*, 24(2):154–160, 2015.
- Eric Mazur. Farewell, lecture. *Science*, 323(5910):50–51, 2009.
- Brian McNely. Backchannel persistence and collaborative meaning-making. In *Proceedings of the 27th ACM international conference on Design of communication*, pages 297–304. ACM, 2009.
- Albert Mehrabian. Silent messages. *PsycINFO Database Record*, 1971.

- Joan Middendorf and Alan Kalish. The “change-up” in lectures. In *Natl. Teach. Learn. Forum*, volume 5, pages 1–5. Wiley Online Library, 1996.
- Adolf Miehle, Ugo Fisch, and Carl-Magnus Eneroth. *Surgery of the facial nerve*. Saunders, 1973.
- Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 37(3):311–324, 2007.
- Michael G Moore. Editorial: Three types of interaction. 1989.
- Kenneth Alberto Funes Mora and Jean-Marc Odobez. Gaze estimation from multimodal kinect data. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 25–30. IEEE, 2012.
- Louis-Philippe Morency, Candace Sidner, Christopher Lee, and Trevor Darrell. Contextual recognition of head gestures. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 18–24. ACM, 2005.
- D. Morris. *Bodytalk: a world guide to gestures*. Jonathan Cape, 1994. ISBN 9780224039697. URL <https://books.google.ch/books?id=MQ5-AAAAMAAJ>.
- Djamel Mostefa, Nicolas Moreau, Khalid Choukri, Gerasimos Potamianos, Stephen M Chu, Ambrish Tyagi, Josep R Casas, Jordi Turmo, Luca Cristoforetti, Francesco Tobia, et al. The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language Resources and Evaluation*, 41(3-4):389–407, 2007.
- Erik Murphy-Chutorian and Mohan M Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626, 2009.
- Laurent Nguyen, Jean-Marc Odobez, and Daniel Gatica-Perez. Using self-context for multimodal detection of head nods in face-to-face interactions. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 289–292. ACM, 2012.
- Marc-Antoine Nüssli. Dual eye-tracking methods for the study of remote collaborative problem solving. 2011.
- Jean-Marc Odobez and Sileye Ba. A cognitive and unsupervised map adaptation approach to the recognition of the focus of attention from head pose. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1379–1382. IEEE, 2007.
- Margarita Osadchy, Yann Le Cun, and Matthew L Miller. Synergistic face detection and pose estimation with energy-based models. *The Journal of Machine Learning Research*, 8:1197–1215, 2007.
- Maja Pantic and Marian Stewart Bartlett. *Machine analysis of facial expressions*. I-Tech Education and Publishing, 2007.

Bibliography

- Maja Pantic, Roderick Cowie, Francesca D'Errico, Dirk Heylen, Marc Mehu, Catherine Pelachaud, Isabella Poggi, Marc Schroeder, and Alessandro Vinciarelli. Social signal processing: the research agenda. In *Visual analysis of humans*, pages 511–538. Springer, 2011.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Alex Pentland. Socially aware, computation and communication. *Computer*, 38(3):33–40, 2005.
- Alex Pentland. To signal is human. *American scientist*, 98(3):204–211, 2010.
- Alex Pentland and Tracy Heibeck. *Honest signals*. MIT press Cambridge, MA, 2008.
- David I Perrett and Nathan J Emery. Understanding the intentions of others from visual signals: Neurophysiological evidence. 1994.
- Rosalind W Picard. *Affective computing*. MIT press, 2000.
- Robert M Pirsig. *Zen and the art of motorcycle maintenance: An inquiry into values*. Random House, 1999.
- Ernst Pöppel and Lewis O Harvey Jr. Light-difference threshold and subjective brightness in the periphery of the visual field. *Psychologische Forschung*, 36(2):145–161, 1973.
- Michael I Posner and Stephen J Boies. Components of attention. *Psychological review*, 78(5): 391, 1971.
- Mirko Raca and Pierre Dillenbourg. System for assessing classroom attention. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 265–269. ACM, 2013.
- Mirko Raca and Pierre Dillenbourg. Holistic analysis of the classroom. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 13–20. ACM, 2014.
- Mirko Raca, Roland Tormey, and Pierre Dillenbourg. Student motion and its potential as a classroom performance metric. In *3rd International Workshop on Teaching Analytics (IWTa)*, number EPFL-TALK-188524, 2013.
- Mirko Raca, Roland Tormey, and Pierre Dillenbourg. Sleepers' lag-study on motion and attention. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, pages 36–43. ACM, 2014.
- Mirko Raca, Lukasz Kidzinski, and Pierre Dillenbourg. Translating head motion into attention - towards processing of student's body-language. In *Proceedings of the Eight International Educational Data Mining*, pages XX–XX. ACM, 2015.

- Bret Range, Heather Duncan, and David Hvidston. How faculty supervise and mentor pre-service teachers: Implications for principal supervision of novice teachers. *International Journal of Educational Leadership Preparation*, 8(2):43–58, 2013.
- David N Rapp. The value of attention aware systems in educational settings. *Computers in Human Behavior*, 22(4):603–614, 2006.
- Siddharth S Rautaray and Anupam Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54, 2015.
- Daniel C Richardson and Rick Dale. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, 29(6):1045–1060, 2005.
- Daniel C Richardson, Rick Dale, and Natasha Z Kirkham. The art of conversation is coordination common ground and the coupling of eye movements during dialogue. *Psychological science*, 18(5):407–413, 2007.
- Virginia P Richmond, James C McCroskey, and Mark Hickson. *Nonverbal behavior in interpersonal relations*. Prentice Hall Englewood Cliffs, NJ, 1991.
- Rutger Rienks, Dong Zhang, Daniel Gatica-Perez, and Wilfried Post. Detection and application of influence rankings in small group meetings. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 257–264. ACM, 2006.
- Maria Rimini-Doering, Dietrich Manstetten, Tobias Altmueller, Ulrich Ladstaetter, and Michael Mahler. Monitoring driver drowsiness and stress in a driving simulator. *Driving Assessment*, 2001.
- Jelena Ristic and Alan Kingstone. Taking control of reflexive social attention. *Cognition*, 94(3): B55–B65, 2005.
- Verónica Rivera-Pelayo, Johannes Munk, Valentin Zacharias, and Simone Braun. Live interest meter: learning from quantified feedback in mass lectures. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 23–27. ACM, 2013.
- Claudia Roda and Julie Thomas. Attention aware systems: Theories, applications, and research agenda. *Computers in Human Behavior*, 22(4):557–587, 2006.
- David Rosengrant, Doug Hearngrington, Kerriann Alvarado, and Danielle Keeble. Following student gaze patterns in physical science lectures. In *AIP Conference Proceedings*, volume 1413, page 323, 2012.
- D.A. Schon. *The reflective practitioner: How professionals think in action*, volume 5126. Basic Books, 1984.
- Lisa M Schreiber, Gregory D Paul, and Lisa R Shibley. The development and test of the public speaking competence rubric. *Communication Education*, 61(3):205–233, 2012.

Bibliography

- Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.
- Kshitij Sharma, Patrick Jermann, and Pierre Dillenbourg. “with-me-ness”: A gaze-measure for students’ attention in moocs. In *International Conference Of The Learning Sciences*, number epfl-conf-201918, 2014.
- David Shenk. *Data smog: Surviving the information glut*. Harper San Francisco, 1998.
- Jianbo Shi and Carlo Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR’94., 1994 IEEE Computer Society Conference on*, pages 593–600. IEEE, 1994.
- Ben Shitrit, J Berclaz, F Fleuret, and P Fua. Multi-commodity network flow for tracking multiple people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.
- Kevin Smith, Sileye O Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Tracking the visual focus of attention for a varying number of wandering people. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(7):1212–1229, 2008.
- Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *Affective Computing, IEEE Transactions on*, 3(2):211–223, 2012.
- Robert Sommer. Studies in personal space. *Sociometry*, 22(3):247–260, 1959.
- ROBERT H Spector. Visual fields. *Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edi. Boston: Butterworths*, pages 565–572, 1990.
- Ellen Stader et al. Expert and novice teachers’ ability to judge student understanding. 1990.
- Rainer Stiefelhagen. Estimating head pose with neural networks-results on the pointing04 icpr workshop evaluation data. In *Pointing’04 ICPR Workshop of the Int. Conf. on Pattern Recognition*, 2004.
- Rainer Stiefelhagen and Jie Zhu. Head orientation and gaze direction in meetings. In *CHI’02 Extended Abstracts on Human Factors in Computing Systems*, pages 858–859. ACM, 2002.
- Rainer Stiefelhagen, Michael Finke, Jie Yang, and Alex Waibel. From gaze to focus of attention. In *Visual Information and Information Systems*, pages 765–772. Springer, 1999.
- Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *Neural Networks, IEEE Transactions on*, 13(4):928–938, 2002.
- Rainer Stiefelhagen, Keni Bernardin, Hazim Kemal Ekenel, J McDonough, Kai Nickel, Michael Voit, and Matthias Wölfel. Audio-visual perception of a lecturer in a smart seminar room. *Signal Processing*, 86(12):3518–3533, 2006.

- Helen Timperley, Aaron Wilson, Heather Barrar, and Irene Fung. Teacher professional learning and development. 2008.
- Dejan Todorović. Geometrical basis of perception of gaze direction. *Vision research*, 46(21): 3549–3562, 2006.
- Roel Vertegaal. The gaze groupware system: mediating joint attention in multiparty communication and collaboration. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 294–301. ACM, 1999.
- Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. Social signals, their function, and automatic analysis: a survey. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 61–68. ACM, 2008.
- Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.
- Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D’Errico, and Marc Schröder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *Affective Computing, IEEE Transactions on*, 3(1):69–87, 2012.
- Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.
- Michael Voit and Rainer Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 173–180. ACM, 2008.
- Michael Voit and Rainer Stiefelhagen. 3d user-perspective, voxel-based estimation of visual focus of attention in dynamic meeting scenarios. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, page 51. ACM, 2010.
- Michael Voit, Kai Nickel, and Rainer Stiefelhagen. Estimating the lecturer’s head pose in seminar scenarios—a multi-view approach. In *Machine Learning for Multimodal Interaction*, pages 230–240. Springer, 2006.
- Michael Voit, Kai Nickel, and Rainer Stiefelhagen. Head pose estimation in single-and multi-view environments-results on the clear’07 benchmarks. In *Multimodal Technologies for Perception of Humans*, pages 307–316. Springer, 2008.
- Alex Waibel, Hartwig Steusloff, Rainer Stiefelhagen, and Kym Watson. *Computers in the human interaction loop*. Springer, 2009.
- Alex Waibell, Hartwig Steusloff, Rainer Stiefelhagen, et al. Chil: Computers in the human interaction loop. 2011.

Bibliography

- Corporation WakaTime. On-line resources. <http://www.wakatime.com>, 2014.
- Nigel Ward and Wataru Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207, 2000.
- Eugene J Webb, Donald Thomas Campbell, Richard D Schwartz, and Lee Sechrest. *Unobtrusive measures: Nonreactive research in the social sciences*, volume 111. Rand McNally Chicago, 1966.
- Michel Wedel and Rik Pieters. A review of eye-tracking research in marketing. *Review of marketing research*, 4(2008):123–147, 2008.
- Mark Weiser. The computer for the 21st century. *Scientific american*, 265(3):94–104, 1991.
- Delores A Westerman. Expert and novice teacher decision making. *Journal of Teacher Education*, 42(4):292–305, 1991.
- Jennie Whitcomb, Hilda Borko, and Dan Liston. Why teach? part ii. *Journal of Teacher Education*, 2008.
- Jacob Whitehill, Marian Bartlett, and Javier Movellan. Automatic facial expression recognition for intelligent tutoring systems. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–6. IEEE, 2008.
- Mani Williams, Jane Burry, Asha Rao, and Nathan Williams. A system for tracking and visualizing social interactions in a collaborative workenvironment. 2015.
- Karen Wilson and James H. Korn. Attention during lectures: Beyond ten minutes. *Teaching of Psychology*, pages 85–89, December 2007.
- Gary Wolf. The quantified self. In *TED conferences web archive*. Presented as the TED at Cannes, 2010.
- Marcelo Worsley and Paulo Blikstein. Towards the development of multimodal action based assessment. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 94–101. ACM, 2013.
- Marcelo Worsley and Paulo Blikstein. Analyzing engineering design through the lens of computation. *Journal of Learning Analytics*, 1(2):151–186, 2014.
- Ying Wu and Kentaro Toyama. Wide-range, person-and illumination-insensitive head orientation estimation. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 183–188. IEEE, 2000.
- Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011.

- Sarita Yardi. The role of the backchannel in collaborative learning environments. In *Proceedings of the 7th international conference on Learning sciences*, pages 852–858. International Society of the Learning Sciences, 2006.
- Xiang Yu, Shaoting Zhang, Zhennan Yan, Fei Yang, Junzhou Huang, Norah E Dunbar, Matthew L Jensen, Judee K Burgoon, and Dimitris N Metaxas. Is interactional dissynchrony a clue to deception? insights from automated analysis of nonverbal visual cues. 2013.
- Zeynep Yücel and Albert Ali Salah. Head pose and neural network based gaze direction estimation for joint attention modeling in embodied agents. In *Proc. Annual Meeting of Cognitive Science Society*, 2009.
- Zeynep Yucel and Albert Ali Salah. Resolution of focus of attention using gaze direction estimation and saliency computation. In *Affective Computing and Intelligent Interaction and Workshops, 2009. AII 2009. 3rd International Conference on*, pages 1–6. IEEE, 2009.
- Zhengyou Zhang. Microsoft kinect sensor and its effect. *MultiMedia, IEEE*, 19(2):4–10, 2012.
- Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.
- Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.

Mirko Raca

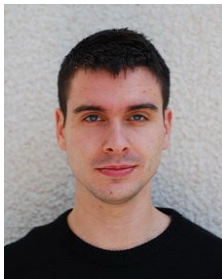
Research Assistant, CHILI laboratory, EPFL

Chemin du Mont-Blanc 11, 1023 Crissier, Switzerland

Language: Serbian (native), English (fluent), French (basic)

(tel) +41 78 905 8782

mirko.raca@epfl.ch ; racamirko@gmail.com



Key points:

- PhD in Computer Vision and statistical data analysis, EPFL
- highly proficient in C++, Python, and quant. data analysis
- 4 years corporate experience, 5 years academic work

Education

- **2010 – present (until October 2015): EPFL (Lausanne, Switzerland) - PhD thesis**, Computer-Human Interaction In Learning And Instruction (CHILI) laboratory, *thesis topic*: Computer-vision based analysis of audience behavior
- **2003 – 2008: Faculty of Technical Science (Novi Sad, Serbia) – Msc**, Applied Computer Science and Informatics, average grade 9.32 (maximum: 10)

Experience

- **2010 – present: EPFL** research assistant – **C++/Python/Java/R/Matlab** algorithm development and analysis of data, mathematical modeling, subjects behavior analysis. Data extraction was implemented in **C++** (UI done with **Qt**, developed on **Linux**), while data analysis were done primarily in **Python** with additional R and Matlab work. Java was used for different data visualizations.
- **2006 – 2010: Schneider Electric / DMS group Serbia** – student intern (2 years) - developer of in-house solutions (**Python, Java**), developer/senior developer (2 years) - outsourced work for Siemens PTD (MS Visual Studio, **C++**, client-server architecture). Responsible for maintaining one of the legacy architectural components for Siemens PTD (Germany) in an international team of 50+ developers. Participated in requirement specification and analysis stages as control.
- **2008: SCA solutions** – UI development of TERA portal (GPS-based vehicle tracking and security). Web-based system for displaying location of vehicles based on Google API, **JavaScript** and **Python**-based Django framework.
- **2005 – 2007: EESTEC Int** – web administrator/maintenance of the existing portal (Linux), web developer of the new **Python**-based portal (plone/zope frameworks).

Skills and interests

- **Technical:** C/C++ (Qt, std, boost, OpenCV, dlib), **Python** (scipy, numpy, pandas, matplotlib), graphics and visualizations (OpenGL/Ogre3D, Processing, NodeBox-gl), Java EE (EJB, Hibernate), Linux (10+ years, desktop and server maintenance), Android, Bash, R, Matlab
- **Research:** applied computer vision techniques (**C++**, face detection/recognition, person tracking, part-based models), statistical data analysis (R and **Python** based tests and analysis), machine learning (supervised, unsupervised)
- **Other skills:** UML specifications, SQL-based database (MySQL, Postgres, Oracle), Android UI prototyping, WebDev (jQuery, JavaScript, Django, zope/plone, HTML, CSS)

Awards

- **2010:** VIP Serbia Android Challenge – honorable mentions – “Happening” application
- **2006, 2007:** Certificate for recognition of achievements, EESTEC Int.
- **2006:** University of Novi Sad - Award for achieved results in the academic year

Selected Publications (all publications available at <http://infoscience.epfl.ch>)

- *Holistic Analysis of the Classroom*. **M. Raca** and P. Dillenbourg. In *Proceedings of the 3rd Multimodal Learning Analytics Workshop and Grand Challenges, ICMI'14*, Istanbul, Turkey, November 12, 2014.
- *Sleepers' lag – study on motion and attention*. **Raca, M.**, Dillenbourg, P., & Tormey, R. In *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (No. EPFL-CONF-196641). ACM. March 24, 2014.
- Translating Head Motion into Attention - Towards Processing of Student's Body-Language. **M. Raca**, L. Kidzinski and P. Dillenbourg. 8th International Conference on Educational Data Mining, Madrid, Spain, June 26-29, 2015.

Academic/Teaching

- Reviewer for JLA (Journal of Learning Analytics), GCMA'15 (Grand Challenge in Multimodal Analytics), ISMAR'14 (International Symposium on Mixed and Augmented Reality)
- Teaching assistant in 6 bachelor-level courses (topics: C++, Java, Object-oriented programming), TA in Coursera-based MOOC (3000+ students)
- Member of the SoLAR (Society for Learning Analytics) and SIG-MLA (Special interest group Multimodal Learning Analytics) since 2014

Personal information

married, guitar player, Kaggle competitor, recreational coder (Project Euler), Github profile (<https://github.com/racamirko>), hiker/mountaineer, marathon runner

Recommendations available upon request

- prof. Pierre Dillenbourg, PhD (EPFL)
- George Srdanov (Verizon Wireless)
- Marjan Povolni (Schneider electric/DMS)