- und wir erarbeiteten eine (allen Auflagen des Datenschutzes gerecht werdende) Form der Anonymisierung der visuellen Daten (= Kantenbilder[2]).

Wegen der genannten Probleme wurden im *sozialwissenschaftlichen* Arbeitsbereich innerhalb des Projektzeitraums, wenn auch mit teils erheblichen Verzögerungen, folgende Daten erhoben und teils hermeneutisch, teils inhaltsanalytisch analysiert: So wurden 42 qualitative, etwa ein- bis zweistündige, themenzentrierte Interviews geführt, größtenteils mit Polizeipersonen aus drei Bundesländern (NRW, Baden-Württemberg und Niedersachsen), die in unterschiedlichen Funktionen wiederholt vor und in Fußballstadien im Einsatz waren (Polizeiführer, Mitglieder von Hundertschaften, Beweis- und Sicherheitsdienst, szenekundige Beamte)[3]. Zudem wurden ein Fanprojektmitarbeiter, drei (Ultra-)Fans und ein Hooligan, eine Person vom Sicherheitspersonal und ein Sportfunktionär ausführlich interviewt. Die Interviews wurden alle inventarisiert (mit kurzer Inhaltsangabe), 37 komplett oder zu großen Teilen transkribiert und inhaltsanalytisch, sowie ausgewählte Stellen mit dem Verfahren der Hermeneutischen Wissenssoziologie in Bezug auf die Projektfragestellung ausgewertet.

Zur *Videoanalyse* wurden Videodaten aus Feldbegehungen, ethnografischen Aufzeichnungen im Fußballstadion (unter Polizeibegleitung), auf Demonstrationen, kirchlichen Veranstaltungen erhoben; es kam zum Datenaustausch mit dem Forschungszentrum Jülich; ebenso wurden zur Projektfragestellung passende Videosequenzen aus öffentlichen Archiven wie YouTube-/Vimeo-/LiveLeak-Videos erhoben und teils sehr ausführlich analysiert – so z. B. eine Sequenz zum Spiel Fortuna Düsseldorf gegen den MSV Duisburg und andererseits die Auseinandersetzung Stuttgarter Ultras mit dem Bielefelder Ordnerpersonal vom 17.04.17. Anhand der Interview-, Feld- und Videodaten, die im Projekt zur Verfügung standen, konnten mittels Datentriangulation und Methodentriangulation einzelne multiperspektivische Rekonstruktionen von Ereignissen vorgenommen werden.

Im Bereich der *Neuroinformatik* wurden erste Ansätze zur automatischen Erfassung von Gruppenemotionen erprobt. Der Ansatz von Choi/Shahid/Sava-

2 Kantenbilder (auch *Gradientenbilder*) sind schwarz-weiß und zeigen nur noch die Umrisse (= Kanten) der gefilmten Personen und Gruppen. Einzelne Personen und deren Gestik und Mimik sind auf Kantenbildern nicht mehr erkennbar. Das stellte für die Analyse dieses Videomaterials eine sehr große Herausforderung dar.

3 Bei diesen Interviews wurde sichtbar, dass nicht alle Landespolizeien die gleichen Einsatz- und Kommunikationsstrategien im Umgang mit Fußballfans haben, sondern dass teils deutliche Unterschiede bestehen. Da es jedoch nicht das Ziel des Projekts war, Unterschiede der landespolizeilichen Behandlung von Eskalationsprozessen zu identifizieren, haben wir in der Auswertung diese Unterschiede zwar zur Kenntnis genommen, jedoch nicht herausgearbeitet, wie sie sich im Einzelnen unterscheiden und welche Folgen dies für das Eskalationsgeschehen hat. Dafür bedürfte es einer gesonderten Studie.

rese (2011) zur *Bestimmung der zeitlichen Entwicklung* der relativen Position und Körperausrichtung einzelner Personen in kleineren Menschengruppen wurde angepasst und anhand zwei verschiedener Datensätze getestet. Aufbauend auf den Arbeiten von Yao/Gall/Van Gool (2010) wurde ein System zur *raumzeitlichen Klassifikation von Handlungen einzelner Personen* in kurzen Videoausschnitten entwickelt und an den o. g. inszenierten Situationen sowie dem öffentlich zugänglichen UCF Sports Action Data Set (Soomro/Zamir 2014; Spata 2016) erprobt. Zur Erkennung anormalen Verhaltens von Menschengruppen wurde der optische Fluss geschätzt. Dabei wurden überall im Bild, wo der lokale Kontrast ausreicht, Bewegungsvektoren geschätzt, ohne zuvor eine Segmentation von Personen oder Objekten zu versuchen. Der gewählte Ansatz baute auf Mehran/Oyama/Shah (2009) auf und leistete eine computational schnelle Schätzung mithilfe einer Particle Advection (Ali/Shah 2007). Das Social Force Model (Helbing/Molnár 1995), das Bewegungstrajektorien von Fußgängern soziale Kräfte innerhalb von Gruppen zuschreibt, wurde dabei unmittelbar auf die Partikel des Schätzalgorithmus angewandt, also ohne diese unbedingt einer Person zuzuschreiben. Schließlich wurde mittels Latent Dirichlet Allocation (Blei/Ng/Jordan 2003) eine Klassifizierung in normales und abnormales Gruppenverhalten vorgenommen (Hegenbarth 2016). Mit der vorliegenden Methodik konnte der plötzliche Umschlag von Verhaltens-/Bewegungsmustern detektiert werden. Um die Dichte von Menschenmengen zu schätzen, wurde für Aufnahmen aus der Totalen ein computational schnelles Verfahren zur Detektion von Köpfen oder Kopf-Schulter-Partien aus dem Zählverfahren von Idrees et al. (2013) abgeleitet. Dieses wurde anhand des öffentlich zugänglichen INRIA Datensatzes (Dalal/Triggs 2005) evaluiert.

4. Probleme der interdisziplinären Projektarbeit[4]

In dem Projekt arbeiteten Soziolog_innen, Kommunikationswissenschaftler_innen, Informatiker_innen und Videoanalytiker_innen zusammen. Die damit einhergehenden interdisziplinären Herausforderungen haben alle Beteiligten im Rückblick klar unterschätzt, trotz einschlägiger Erfahrungen. Neben praktischen Problemen gab es eine Reihe von inhaltlichen und theoretischen Missverständnissen, welche die Arbeit immer wieder erschwerten. Manchmal erkannten wir erst nach Wochen, dass wir uns missverstanden hatten. Dies wurde im Verlauf der Projektarbeit deutlich und äußerte sich in unterschiedlichen Schwerpunktsetzungen und Planungen des Zeitverlaufs der Projektarbeit sowie auch in un-

4 Den im folgenden Kapitel benannten Problemen liegt ein Papier von Gregor Schöner zugrunde, das ich hier ausführlich wiedergebe und ergänze.

erkannten Unterschieden bei der Nutzung von Begriffen wie Varianz und bei den der Forschung zugrunde gelegten Begriffe wie *Emotion*, *Eskalation*, *Gewalt*. Aber auch die Begriffe *Gruppe*, *Masse* und *Ansammlung* waren nicht hinreichend geklärt, um für die Analyse nützlich zu sein.

Gleiches stellte sich auch bei dem Phänomen der ‚Übertragung von Emotionen' heraus. Die hierzu in der Literatur vorliegenden Theorien benutzen vor allem die Metapher der Ansteckung, der Resonanz, des Schwarmverhaltens oder der unbewussten Kommunikation.

Der Ansatz des technischen Vorgehens war es, Algorithmen des maschinellen Lernens zu nutzen, um die Detektion und Klassifikation von Ereignissen, die verschiedene Gruppenemotionen signalisieren, zu automatisieren. Solche Algorithmen lernen aus Beispielen, hier also aus Videomaterial, in dem menschliche Beobachter_innen Ereignisse gekennzeichnet haben (‚labeln'). Je größer die Vielfalt der visuellen Erscheinung der gleichen Klasse, umso mehr Beispiele sind notwendig. Durch Vorverarbeitung der Bilder kann ein Teil der Bildvarianz eliminiert werden, was die Lernaufgabe erleichtert.

Die Aufgabenverteilung im interdisziplinären Team schien offensichtlich. Die Techniker_innen beschäftigen sich mit Algorithmen der Vorverarbeitung und dem Entwurf der spezifischen Algorithmen aus dem Bereich maschinelles Lernen. Die Soziolog_innen liefern Beispieldaten als gelabelte Videoabschnitte und damit implizit die grundlegenden Begriffe zur Beschreibung von Gruppenemotionen. Im Verlaufe des Projekts zeigte sich, dass diese Schnittstelle zwischen Technik und menschlichem Beobachten bei weitem nicht genügend spezifiziert war. Im Grunde hatten beide Seiten eine ganze Reihe von impliziten Annahmen und Randbedingungen nicht kommuniziert und waren sich dieser Unterlassung zunächst gar nicht bewusst.

Eine Frage ist, ob gelabelte Ereignisse lokalisiert im Bild oder global der Gesamtsituation zugeordnet sind. Die Verfahren des maschinellen Lernens profitieren stark von lokalisierten Labeln, denn diese erlauben, die gesamte Varianz des Bildmaterials außerhalb der lokalisierten ‚region of interest' (ROI) zu eliminieren. Die Soziolog_innen gingen davon aus, dass Labels lokal in der Zeit sind, aber global die gesamte visuelle Szene bezeichnen. Das Software-Instrument der Soziolog_innen sah auch gar keine Lokalisierung der zuzuweisenden Label vor.

Die Methoden des maschinellen Lernens zielen zunächst auf eine einfache ‚Ein-Klassen-Klassifikation' ab, in dem Bilder oder ROI also als zugehörig oder nicht zugehörig zu einer bestimmten Klasse erkannt werden (‚binäre Klassifikation'). Mehrklassen-Klassifikation ist möglich, verlangt aber entsprechend größere Mengen von Beispielen. Die implizite Erwartung der maschinellen Lernern war folglich, dass die Soziolog_innen einige wenige, grundlegende Klassen vorschlagen, für die sie dann viele Beispiele liefern.

Für die Soziolog_innen war dagegen die Differenzierung der beschreibenden Begriffe ein wesentliches Ziel. Sie versuchten für jedes einzelne Ereignis

eine Charakterisierung zu erreichen, die nicht nur dem oberflächlich sichtbaren, sondern auch den mitempfundenen, vom Kontext abhängigen Dimensionen des Geschehens Rechnung trug. Eine Vielzahl beschreibender Begriffe wurde erarbeitet, und es war gerade diese Erarbeitung, die einen wichtigen Teil des Erkenntnisprozesses ausmachte. Viele Beispiele für die genau gleiche kategorielle Beschreibung zu liefern, lief diesem Bestreben entgegen.

Erschwerend für die interdisziplinäre Zusammenarbeit war in diesem Kontext auch die unterschiedliche Auffassung vom eigentlichen Erkenntnisprozess. Für die Soziolog_innen war die schrittweise Aufdeckung von möglichen Ereignissen, deren Charakterisierung und Differenzierung eine natürliche Form des Vorgehens. Für das maschinelle Lernen konnte die Arbeit erst beginnen, wenn konkrete Hypothesen in Form von gelabelten Beispielen in einem Umfang vorlagen, der ermöglichte, auf einer Teilmenge maschinell zu lernen, um die übrigen Daten zum Test der Generalisierung zu nutzen. Folglich war auch die zeitliche Koordination der Zusammenarbeit nicht einfach zu bewerkstelligen.

Neben diesen recht unterschiedlichen grundsätzlichen Perspektiven gab es kleinere Unterschiede. So war etwa für die Soziolog_innen eine bewegte Kamera, die ins Geschehen integriert ist, unter Umständen ausdrucksstärker als eine statische Kamera, die das Geschehen von außen registriert. Für die maschinellen Lerner ist dagegen eine Eigenbewegung der Kamera eine Herausforderung, da Bewegung im Bild ein wirksames Mittel zur Vorauswahl von Bildregionen ist, in denen Ereignisse auftreten mögen. Bei bewegter Kamera verliert dieser Salienzkanal seine Wirkung.

Auch die Frage der multisensoriellen Basis mancher Begriffe, insbesondere zur Hinzunahme von auditorischen Hinweisen, wurde lange diskutiert. Für die Soziolog_innen sind dies natürliche Dimensionen des zu verstehenden Geschehens. Für die Bildverarbeitung entsteht bei der Miteinbeziehung multisensorieller Hinweise die Notwendigkeit, zusätzliche Disziplinen, wie die automatische Analyse akustischer Kanäle, zu beachten. Dabei würde gleichzeitig die potenzielle Varianz des Datenmaterials weiter erhöht und entsprechend würden sich die Anforderungen an die gelabelten Beispieldaten weiter erhöhen.

Diese Varianz innerhalb des Videomaterials, das der automatischen Detektion von Gruppenemotionen dienen sollte, war ein Thema, dessen Bedeutung wir erst im Laufe des Projektes für beide Disziplinen klar artikulieren konnten. Aus Sicht der maschinellen Lerner geht es um die Varianz auf Bildebene, also die Variabilität der visuellen Erscheinung der relevanten Ereignisse, die in verschiedenen Beispielen auftritt. Wird, beispielsweise, die Aktivität, die den Fokus einer aufkeimenden Störung darstellt, im Bild innerhalb einer kleinen Region erfasst, weil das Geschehen entsprechend weit entfernt ist, so lernt ein Algorithmus diese visuelle Erscheinungsform mit. Will man das entsprechende Ereignis unter anderen Umständen, wenn der Fokus näher an der Kamera liegt, erkennen (und somit generalisieren), so muss das Lernmaterial diese Varianz

der ‚region of interest' in ihrer Größe im Bild enthalten, also Beispiele von visuell kleinen, mittleren, und großen Abbildungen der Erscheinung ‚Fokus' liefern. Viele andere Dimensionen der visuellen Erscheinung von Ereignissen sind nicht so einfach zu umschreiben wie die Größe im Bild, sodass es auch keine einfache Art gibt, diese Form von Varianz durch theoretische Überlegungen zu eliminieren. Diese Form der Varianz ist also zu unterscheiden von der Varianz, welche die Soziolog_innen durchaus interessiert und die ja gerade der Tendenz zugrunde liegt, Ereignisse durch eine Vielzahl von Begriffen differenziert zu beschreiben.

Die zum maschinellen Lernen nutzbaren Beispielvideos waren jedoch radikal eingeschränkt durch Fragen der Filmqualität (bewegte Kamera, wechselnder Zoom, wechselnde Bildausschnitte) und durch die darin enthaltene visuelle Varianz (Hintergrund, Dichte, Abstand von der Szene, Szenenelemente). Die selbst produzierten Videos von einfach strukturierten Eskalationen konnten einerseits die Komplexität des Geschehens nicht erfassen, lieferten aber andererseits immer noch viel zu wenig Lernbeispiele um maschinelle Lernmethoden einzusetzen.

In der Begrifflichkeit der maschinellen Lerner litt das Projekt letztlich also an einer Knappheit von Daten im Sinne von einer hinreichend großen Anzahl von visuellen Beispielen für eine kleine Anzahl von Labeln. Für die Soziolog_innen war die Aufgabe, die vorhandenen Daten in ihrer Begrifflichkeit zu analysieren, schon sehr umfangreich und die Forderung nach immer mehr ‚Daten' schwer verständlich. Neben den geschilderten begrifflichen Schwierigkeiten war auch der in Deutschland sehr rigoros praktizierte Datenschutz ein Begrenzungsfaktor. Die Zusammenarbeit mit Veranstaltern wurde dadurch ebenso stark beeinträchtigt wie die Zusammenarbeit mit den im Grunde sehr interessierten polizeilichen Stellen.

Der Dialog über alle diese Probleme zwischen den Forscher_innen der unterschiedlichen Disziplinen nahm verschiedene Formen an. Die maschinellen Lerner implementierten beispielhafte Algorithmen und stützten sich dabei auf öffentlichen Datenbasen (meist aus dem amerikanischen Raum). Diese sollte zeigen, was man ‚im Prinzip' tun könnte. Die Soziolog_innen haben die maschinellen Lerner zu Workshops mitgenommen, bei denen die soziologische Videoanalyse trainiert wurde.

Ein Ergebnis dieses intensiven Austausches war die Kristallisierung der Frage, ob Phänomene im Bereich der Gruppenemotionen notwendig zuerst die Segmentierung im Bild der Einzelperson und die Entdeckung ihrer Handlungsintention erfordert. Aus soziologischer Sicht hat diese Annahme tiefe Gründe in der Handlungstheorie. Aus Sicht des maschinellen Lernens löst man bei diesem Vorgehen sehr schwierige Probleme des Computersehens, Segmentation und Intentionserkennung als ersten Schritt in einem Prozess, der dann auf recht einfache Klassifikationen der Gruppenemotion hinausläuft. So kam die