2023 European Conference on Computing in Construction
40th International CIB W78 Conference
Heraklion, Crete, Greece
July 10-12, 2023

CONCRETE FLOW TRANSFORMER: PREDICTING FRESH CONCRETE PROPERTIES FROM CONCRETE FLOW USING VISION TRANSFORMERS

Max Coenen, Christian Vogel, Tobias Schack, Michael Haist
Institute of Building Materials Science, Leibniz University Hannover, Germany

Abstract

To improve sustainability, concretes are increasingly produced using recipes containing up to a dozen different raw materials. The increasing complexity of the composition leads to an increased sensitivity and decreased robustness of the concrete, making a reliable quality control of the concrete highly important. Despite that, current quality control is mainly conducted based on analogous and empirical tests. This paper presents a novel approach for an automatic quality assessment of fresh concrete on the construction site. Based on a camera sensor setup, delivering image sequences showing the concrete flow during the discharge process of a mixing truck, we propose the Concrete Flow Transformer, a deep learning approach based on Vision Transformers, for the prediction of fresh concrete properties. The performance of the proposed approach is evaluated on a challenging real-world data set, demonstrating highly convincing results for the prediction of both, the consistency and rheological parameters of the fresh concrete.

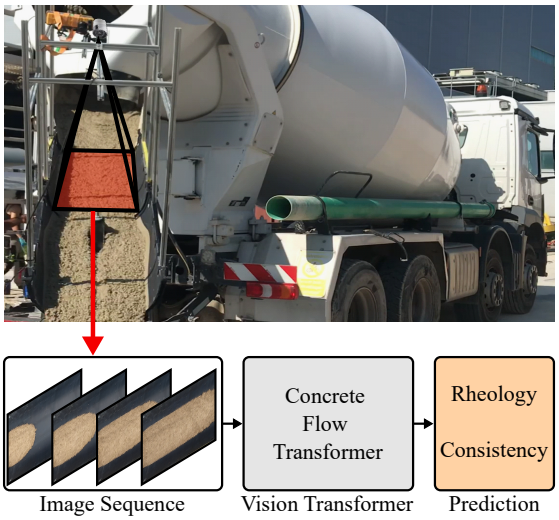


Figure 1: High level overview on our approach. Image sequences are recorded at the concrete discharge channel of a mixing truck. The proposed Concrete Flow Transformer is used to predict the consistency and the rheological properties of the fresh concrete.

Introduction

Fresh concrete can be characterised by its *workability*, a term which describes the concrete's *consistency* and its *rheological* properties. The workability of concrete is a highly important factor for both, the classical casting of concrete at the construction site as well as for 3D printing. In this context, fresh concrete inheriting unsuitable properties can lead to serious quality and safety relevant problems, like segregation, flow blockage, or substantial voids inside the concrete structure; effects demanding for a proper quality control of the fresh concrete before its incorporation. Moreover, in order to meet sustainability goals, concretes are nowadays increasingly produced using recipes containing up to a dozen different raw material components, including e.g. CO_2 reduced cements or recycled aggregates. These increasingly complex mixtures, however, lead to a pronounced sensitivity of the concrete to fluctuations in the raw material properties, and, therefore, to a decreased robustness of the concrete, rendering a thorough quality control even more important today and in the near future.

In current practice, quality control is typically conducted on site based on normative regulated test methods (e.g. flow table test and slump test), which, however, rely on very simple empirical procedures, delivering only provisional information on the concrete's consistency. Rheometer tests on the other hand allow deeper insights into the rheological properties of concrete, but are typically expensive, time consuming, and hardly applicable to standard concrete, rendering them unsuitable for an application in construction practice.

In order to overcome current limitations and to improve concrete quality control and safety assurance on construction sites, we present a novel test procedure for the automatic on-site characterisation of fresh concrete during the discharge process of a mixing vehicle (cf. Fig. 1). As contributions of this paper, we propose a computer vision based strategy for a digital characterisation of fresh concrete based on image sequences showing the concrete flow at the mixing trucks's discharge. Given the recorded video frames, we propose *Concrete Flow Transformer*, a deep learning approach based on Vision Transformers (ViT), for the prediction of concrete properties like consistency and rheological parameters. In this context, we present *flow tokenisation*, a strategy for the generation of patch embeddings serving as input to the ViT, using priorly computed dense optical flow information. Finally, we evaluate the performance of our approach on a challenging real-world data set, demonstrating highly promising results.

We hope with this paper to initiate and encourage further future research, bridging the scientific disciplines of civil engineering, building materials, computer vision, and data sciences in order to bring forth developments of novel, data-driven methods for an automatic, digital, and improved quality assurance in construction and civil engineering.

State of the art

Despite the increasing developments in digitisation and automation of processes in other manufacturing industries, the quality control of the concrete production and construction industry is still based on conventional, non-digital, batch-based and primarily manual methods. Although of essential importance, the testing of concrete properties on the construction site before casting is based on very simple and empirical test methods as e.g. the flow table test (EN 12350-5, 2019) for deriving the concrete's workability. In this test, the fresh concrete is spread on a flow table and the diameter δ [mm] of the resulting slump cake is used to derive an assessment of the concrete's consistency class C. In addition to slump testing, rheometer test methods have gained attraction in testing of fresh concrete (Haist et al., 2020) by determining the parameters of a Bingham model (Yahia et al., 2016) which is used to describe the rheological properties of fresh concrete by the plastic viscosity μ [Pa·s] and the yield stress τ_0 [Pa]. However, concrete rheometer test are exclusively batch-based, laborious and the data interpretation is highly challenging. As a consequence, increasing interest has emerged in developing and providing digital and automatic methods for quality control in the concrete production industry (Haist et al., 2022).

A first approach for fresh concrete monitoring has been proposed in (Yang et al., 2020), where the concrete mix proportion is determined from single images of fresh concrete using a convolutional neural network (CNN). However, in practice, not only the mix proportion but rather an indication for the concrete's workability is of main interest. In (Coenen et al., 2022), an approach for the panoptic segmentation of fresh concrete was presented, which allows to derive conclusions about the sedimentation stability of the concrete, but does not give indications about the concrete's workability. Ponick et al. (2022) proposed an approach for predicting the rheology of concrete from stereo-images and 3D reconstructions of the concrete's surface during the mixing procedure. However, in their method, the temporal information, namely the concrete flow, is not considered. Yet, we believe, that the flow behaviour of fresh concrete carries valuable information on the concrete's characteristics. In (Ding and An, 2018), an approach for determining the workability from image sequences acquired during the mixing process using a LSTM deep learning network has been proposed. While promising results were obtained, processing was done on rather low resolution grey scale images only and the approach relied on 2D transformations, ignoring the clearly visible effects perspective distortions. In this paper, we follow a similar idea, namely to determine the fresh concrete's characteristics from image sequences observing the concrete flow in an open-channel geometry.

From a methodological point of view, the problem tackled in this paper is closely related to deep learning based video interpretation and classification. Compared to single image interpretation, image sequences contain additional temporal information, like e.g. object motion, which is expected to carry valuable and relevant cues for solving the respective problem. Many approaches apply a CNN to each individual video frame in order to leverage the strength of single-image CNNs and aggregate the extracted information across time in order to capture temporal relations. Often, a 2D-CNN backbone is applied to extract frame-wise feature representations and temporal relationships are modelled e.g. as conditional random fields (CRF) (Sigurdsson et al., 2017), via recurrent modules like long-short term memory cells (Hochreiter and Schmidhuber, 1997; Donahue et al., 2015), or via the self-attention mechanism within transformer-based architectures (Zhou et al., 2018; Wang et al., 2021).

By expanding the 2D filters of a CNN to three dimensions, an approach called 3D CNN (Ji et al., 2012), and applying the 3D filters along the temporal domain of an image sequence in addition to the spatial image domain, single-frame CNNs can be generalised to an application to video data. In contrast to 2D CNNs, 3D filters conceptually allow CNNs to model motion because they act as local spatio-temporal filters, thus generating spatio-temporal feature maps. In this way, 3D CNNs are often applied to sequential image data in the literature in order to solve video interpretation problems, cf. for instance (Ji et al., 2012; Tran et al., 2015; Camgoz et al., 2016).

Another strategy that can be found in the literature for video interpretation is to make use of two-stream convolutional neural networks (Simonyan and Zisserman, 2014). In this case, the network architecture is based on two separate recognition streams, a spatial and a temporal stream, which are then combined by fusion at a later stage of the network. While the spatial stream performs video interpretation from still video frames, the temporal stream uses pre-computed dense optical flow maps as input and is trained to generate feature embeddings from the explicit motion information (Wang et al., 2018; Feichtenhofer et al., 2016). In this work, we also make use of explicit optical flow computations, i.e. of explicit per-pixel motion estimates. In addition to the valuable information the optical flow carries, it is also invariant to e.g. texture, colour, and illumination, making it less prone to overfitting effects.

Methodology

The method proposed in this paper is built on the premise that fresh concretes with different rheological properties exhibit a different and distinguishable flow behaviour. This hypothesis is founded by finite-element formulations of non-Newtonian fluid flow, on the basis of which the flow behaviour of Bingham fluid's (such as concrete) can be described as a function of its rheological parameters, namely the plastic viscosity μ and yield stress τ_0 (Whipple, 1997). This paper targets the inverse problem, namely the objective of predicting the rheological properties (τ_0 and μ) and the consistency of fresh concrete from observations of its

flow behaviour. More specifically, we make use of image sequences showing the concrete flow during the discharge process of a mixing vehicle as input data to our approach. The image sequences are acquired using a rigid camera setup installed at the outlet of the mixing vehicle, where we assume the images to be approximately acquired in nadir view to the discharge channel, and the image coordinate axes (x and y) to be approximately axis-parallel to the longitudinal and transversal axes of the channel, resulting in the property that the major constituent of the flow movement is observed along one of the respective axes (here: the x-axis).

Formal problem definition

Given an image sequence $I(\mathbf{t})$ containing the images $I(t_i)$ acquired according to the setup described above and showing the discharge procedure of fresh concrete at discrete time steps $t_i \in \mathbf{t}$ with $\mathbf{t} = [t_0, t_n]$, the goal is to automatically derive the target parameters describing the rheology and consistency of the fresh concrete. For notation, we associate each concrete with its state vector $\mathbf{s} = (\mu, \tau_0, \delta, C)$ comprising the rheology (Bingham yield stress τ_0 [Pa] and plastic viscosity μ [Pa·s]) and the consistency (slump flow diameter δ [mm] and consistency class C). In a first step, we compute the dense optical flow $O(t_i)$ for each image frame t_i, describing the pixelwise movement between the respective image and its subsequent frame. The dense flow maps are used to generate a sequence of flow tokens $\mathbf{z}(\mathbf{t})$ which serves as input to a multi-task Vision Transformer (ViT), that maps the observed concrete flow of the time span $\mathbf{t}$, represented by the flow token sequence, to the corresponding fresh concrete properties $\mathbf{s}$. An overview on the procedure is shown in Fig. 3 and the approach is described in more detail in the following sections.

Dense optical flow

Dense optical flow (cf. Fig. 2) denotes the problem of per-pixel motion estimation between two consecutive frames t_i and t_{i+1} of an image sequence, and, thus, implies the computation of a translation vector $\Delta = (\Delta x, \Delta y)$ for each pixel, describing the pixel's displacement between the two frames in the x and y coordinate direction, respectively.

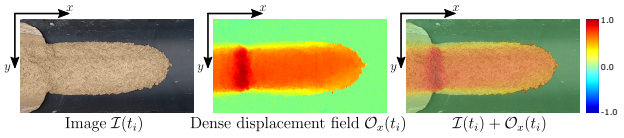


*Figure 2: Visualisation of an example of the dense optical flow computation. **Left**: Original image. **Centre**: The dense displacement field in the x-coordinate direction (colour coding corresponds to the normalised magnitude of the pixelwise displacement Δ_x). **Right**: The RGB image overlayed with the dense displacement field.*

In this paper, we make use of the approach by Farnebäck (2003) in order to derive the two-dimensional dense optical flow field $O(t_i) = (O_x(t_i),\ O_y(t_i))$ for each image $I(t_i)$, whereas $O_x(t_i)$ and $O_y(t_i)$ contain the displacement of each pixel in the x and y coordinate direction, respectively. Regarding the application addressed in this paper, where the open-channel concrete flow is observed, the dense optical flow map implicitly encodes the velocity ϑ of the concrete flow at each pixel position in $[\mathrm{px}/(t_{i+1} - t_i)]$, which corresponds to the magnitude of the translation vector Δ. Furthermore, the flow direction of the concrete, namely the angle α of the translation vector Δ, is also implicitly contained. Fig. 2 shows an example of the dense optical flow map O_x for the concrete flow of a specific epoch t_i.

Revisiting Vision Transformers (ViT)

The original Vision Transformer (ViT) (Dosovitskiy et al., 2021) takes a single image as input and extracts n non-overlapping image patches $x_i \in \mathbb{R}^{h \times w}$ which are transformed into 1D tokens $z_i \in \mathbb{R}^d$ of length d using a linear projection $\mathbf{E}$. The sequence of tokens $\mathbf{z}^0 \in \mathbb{R}^{(n+1) \times d}$ with

$$\mathbf{z}^0 = [z_{\mathrm{cls}}, \mathbf{E}z_1, \mathbf{E}z_2, ..., \mathbf{E}z_n] + \mathbf{p} \tag{1}$$

then serves as input to a transformer encoder architecture (Vaswani et al., 2017). As is shown in Eq. 1, a learnable classification token z_{cls} is prepended to the sequence, whose representation at the final layer of the encoder is used as input embedding for the output layer. Furthermore, a learnable position embedding $\mathbf{p} \in \mathbb{R}^{n \times d}$ is added to the tokens (cf. Eq. 1) in order to retain positional information throughout the permutation invariant self-attention operations of the encoder. The tokens are passed through the transformer encoder which consists of a stack of $l = 1...L$ residual layers, each comprising Multi-Head Self-Attention (MSA) (Vaswani et al., 2017), layer normalisation (LN), and Multi-Layer Perceptron (MLP) blocks, producing the intermediate outputs $\mathbf{z}^l$ with

$$\mathbf{y}^l = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}^{l-1})) + \mathbf{z}^{l-1} \tag{2}$$

$$\mathbf{z}^l = \mathrm{MLP}(\mathrm{LN}(\mathbf{y}^l)) + \mathbf{y}^l. \tag{3}$$

Here, The MLP blocks consist of two linear projections separated by a GELU non-linearity. MSA is based on the self-attention (SA) mechanism, whose goal is to capture the interaction and dependencies amongst all entities (tokens). To this end, three learnable weight matrices $W^Q \in \mathbb{R}^{d \times d_q}$, $W^K \in \mathbb{R}^{d \times d_k}$, and $W^V \in \mathbb{R}^{d \times d_v}$ are defined (where $d_q = d_v$) in order to compute Queries $Q = \mathbf{z}W^Q$, Keys $K = \mathbf{z}W^K$, and Values $V = \mathbf{z}W^V$. The self-attention layer output results to

$$\mathrm{SA} = \mathrm{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_q}}\right) \cdot V. \tag{4}$$

MSA extends the self-attention mechanism by concatenating the outputs of h separate SA heads and projecting them to the final embedding using another learnable weight matrix W^M, such that

$$MSA = \mathrm{Concat}(\mathrm{SA}_1, \mathrm{SA}_2, ..., \mathrm{SA}_h) \cdot W^M. \tag{5}$$

Finally, a MLP head is used on top of the transformer encoder to produce the prediction output based on the final encoded class token embedding z^L_{cls}.

Concrete Flow Transformer

In this section, we describe the *Concrete Flow Transformer*, a ViT-based approach for the prediction of fresh concrete properties from concrete flow observations. An overview on the workflow of the proposed approach is shown in Fig. 3. In contrast to the description in the

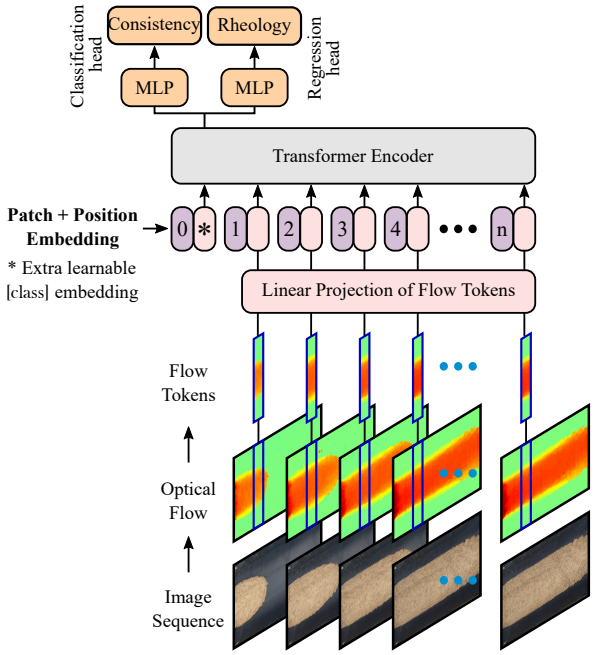


Figure 3: Overview on the procedure of the proposed Concrete Flow Transformer. Given an image sequence, the dense optical flow is computed for each frame. For each flow map, flow tokens are extracted, linearly projected, enriched by position and class embeddings, and fed to a multi-task transformer encoder. Two network heads predict the target parameters describing the consistency and rheology of the fresh concrete.

section above, the approach presented in this paper does not operate on a single image, but on a sequence of images showing the open channel flow of fresh concrete, instead. However, the transformer encoder (Vaswani et al., 2017), which forms the basis of ViT (Dosovitskiy et al., 2021), is a flexible architecture that can operate on any sequence of input tokens $\mathbf{z} \in \mathbb{R}^{(n+1)\times d}$. In this paper, we therefore propose *flow tokenisation*, a strategy for tokenising the input image sequences containing the flow information of the concrete.

Flow tokenisation: As input, we make use of the dense optical flow maps $O(\mathbf{t})$ computed from the image sequence $I(\mathbf{t})$ for each epoch $t_i = t_0 ... t_n$, carrying the motion information representing the concrete flow behaviour over time. The specific setting treated in this work in form of the image sequence acquired from the open channel concrete flow leads to the occurrence of large amounts of redundant information in the data. This is caused by the fact, that the concrete's movement mainly takes place along one direction with a motion behaviour (velocity) that is approximately constant along the direction of movement for an individual time step t_i (cf. Fig. 2). We therefore define a location x_S on the x-coordinate axis of the optical flow map and extract a vertical profile slice $z_x(t_i) \in \mathbb{R}^{y\times 1}$ and $z_y(t_i) \in \mathbb{R}^{y\times 1}$ with a width of 1 [px] at this location from each of both channels of the dense flow field $O_x(t_i)$ and $O_y(t_i)$. We linearly project the two-channel profile slices in order to receive a flow token $z_i \in \mathbb{R}^d$ with $d = 2y$ for each epoch. The tokens of all epochs $t_i \in \mathbf{t}$ with $i = 0...n$ are arranged to the sequence $\mathbf{z}(\mathbf{t}) \in \mathbb{R}^{(n+1)\times d}$ according to Eq. 1 and serve as input to the transformer encoder. A schematic overview of this procedure is shown in Fig. 3.

We argue that by slicing the flow fields in order to generate the flow tokens, we discard the redundant information contained in the flow data while conserving the relevant motion information. As a result from this, the required computational complexity is reduced while, at the same time, the learning complexity of the transformer built on top of the flow tokens is simplified, since in this way, learning to discern relevant data from redundant information becomes less demanding. Particularly, the proposed way of *flow tokenisation* preserves the following features. Each token in $\mathbf{z}(\mathbf{t})$ carries the information of the concrete's flow velocity and flow direction at a specific point in time. As established in (Whipple, 1997), the velocity is a function of a material's rheological properties and, consequently, constitutes one of the most significant cues for the characterisation of the fresh concrete.

Furthermore, the token sequence contains the flow data over the entire time span $\mathbf{t}$, consequently giving the transformer access to information on changes in the velocity profiles over time. These changes correspond to acceleration and deceleration behaviour of the material and can be caused by varying transportation rates of the mixing vehicle. We argue, that velocity changes in the concrete flow as reaction to variations in the transportation rates carry additional valuable information encoding rheological properties of the material.

At last, each token preserves the information of the flow behaviour along the transversal (y) direction of the open channel. Consequently, each token encodes the spatial motion information such as e.g. the current width of the concrete flow stream, which implicitly contains information about the volumetric transportation rate of the concrete.

Prediction: As outlined before, we aim at learning a ViT which predicts the fresh concrete properties represented by the state vector $\mathbf{s}$ from open channel concrete flow observations over a time interval $\mathbf{t}$ represented by the flow tokens $\mathbf{z}(\mathbf{t})$. Towards this goal, we employ a transformer encoder with a number of $\mathbf{L}$ transformer blocks (cf. description above), which takes the sequence of n flow tokens as input and produces a sequence of feature embeddings $\mathbf{z}^L = [z^L_{\mathrm{cls}}, z^L_1, z^L_2, \ldots, z^L_{\mathrm{n}}]$ as output of the final layer according to Eq. 3. As is depicted in Fig. 3, we employ two MLP heads, a *regression head* and a *classification head*, on top of the transformer encoder which predict the target parameters from the class token embedding z^L_{cls} .

The regression head is used to predict a number of n_{Reg} real-valued parameters. In this work, $n_{\text{Reg}} = 3$, where each parameter is related to the concrete's rheology and consistency, namely the slump flow diameter δ, the plastic viscosity μ, and the yield stress τ_0. In this paper, we consider these parameters to be normalised in the range [0; 1] using the minimum and maximum valid values. In that case, we make use of an MLP head with three output variables, one for each parameter, and by using the sigmoid function as activation.

The classification head directly delivers a prediction of the concrete's consistency class C. To this end, the final output is produced by an MLP using a softmax activation function. More specifically, the classification head produces n_{Class} output variables p with $\sum_{j=1}^{N} p_j = 1$ where each variable p_j represents the predicted probability for each respective consistency class C_j. The final class being considered as the predicted class for the input is defined based on the maximum a posteriori criteria, i.e. is chosen according to the class receiving the maximum probability by the ViT.

Training: Training in this work is performed in a supervised manner. Therefore, the training procedure requires training samples in the form of the flow token sequences including corresponding reference values for the concrete's properties. Starting from a random initialisation of the ViT parameters, training is done by minimising a loss function $\mathscr{L}$. The optimisation is performed iteratively using stochastic mini-batch gradient descent (SGD) (Goodfellow et al., 2016), where in this work the *Adam* optimiser (Kingma and Ba, 2015) is used for weight optimisation. The loss function used in this paper is composed of the regression loss $\mathscr{L}_R$ computed for the output of the regression head and the classification loss $\mathscr{L}_C$ computed for the output of the classification head, such that $\mathscr{L} = \mathscr{L}_R + \mathscr{L}_C$. Both losses quantify the difference between the concrete properties predicted by the network and the reference properties. In case of the regression head, the mean squared error (MSE) is used to calculate the loss while the categorical cross-entropy (CE) is used as loss function for the classification head.

Experimental setup

Test data: The empirical evaluation of the proposed method is conducted on a self-acquired data set of open-channel concrete flow recorded during the discharge process of a mixing vehicle. For data acquisition a setup as shown in Fig. 4, consisting of a semi-circular open channel and a camera rig in approximate nadir view, was used. Image acquisition was conducted using a frame rate of 60 [fps]. For the experiments in this paper, we down-scaled the images to a resolution of 380×675 [px], corresponding to an image scale of approximately 1 $[px/mm]$. In total, the acquired data set consists of recordings of the discharge of a variety of 18 different concretes with strongly varying compositions and properties.

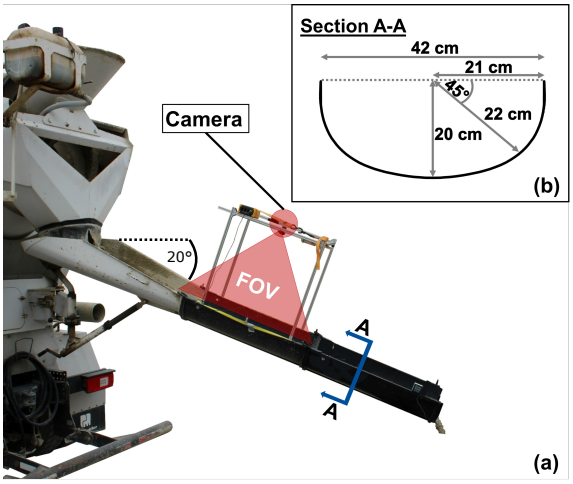


Figure 4: The proposed acquisition setup for the open-channel concrete flow at the discharge of a mixing vehicle.

For this paper, we created sequence snippets, each of which has a length of 60 [s]≙3600 frames, extracted from the original sequence of the 18 concretes between points in time where a representative concrete flow was present and observed during the discharge. In order to obtain groundtruth values for the consistency (slump flow δ and consistency class C) as well as for the rheological parameters (plastic viscosity μ and yield stress τ_0), reference measurements were conducted based on samples taken from each of the 18 concretes. In this context, a portable rheometer for fresh concrete, an eBT-V rheometer (Schleibinger), was used to generate rotational-controlled rheological measurements from which the rheological parameters are derived using the Bingham model according to (Haist et al., 2020). In order to obtain reference values for the consistency, the slump flow was determined according to EN 206-1 (2001)/DIN 1045-2 (2008).

Training: For the experiments in this work, we make use of the *ViT-Base* architecture as defined in (Dosovitskiy et al., 2021) as encoder backbone, consisting of $L = 12$ layers and $h = 12$ MSA heads. Data-wise, we defined both, a spatially and temporally disjunct separation of train, test, and validation splits from the given sequence snippets. For a spatial separation of the data, we defined five equidistant slice locations $x_{S,1}$ to $x_{S,5}$ over the x-axis range of the images, from which the flow tokens are extracted. Three of the five splits are used for training and one is used for validation and testing, respectively. For a temporal separation, we divided the sequence snippets into four equally sized parts, two of them are used for training, and two for validation and testing. Provided the data splits as just described, training is done based on a randomly selected subset of flow tokens extracted from the data. In this paper, we choose $n = 256$ for the number of tokens in each sequence presented to the transformer, meaning that only the flow information of

$256/60 \approx 4$[sec] (supposing a frame rate of 60 fps) are fed as training information to the network for each sample. Training is done using the Adam optimiser (Kingma and Ba, 2015), a mini-batch size of 32 and an initial learning rate if 10^{-4}. To improve training, the learning rate is decreased by a factor of 10^{-1} after 50 epochs with no improvement in the training loss.

Evaluation strategy: For the empirical evaluation of the proposed approach, we apply a sliding window strategy using the same number of flow tokens as used for training. More specifically, for each test split of the different concrete snippets of length $\mathbf{t}$, we extract a token sequence of size $n = 256$ covering the epochs t_i to t_{i+256}, and incrementally increase i by 1. Each token sequence is processed individually by the ViT, resulting in a number of $\mathbf{t} - 256$ predictions for each of the concrete's state parameters $\mathbf{s} = (\mu, \tau_0, \delta, C)$. In order to assess the performance of the classification head, namely the prediction of the consistency class C, we determine the confusion matrix of the predictions. Furthermore, we compute values for the overall accuracy (OA) of the predictions as well as class-wise values for *recall*, *precision*, and *F1-score*. For the evaluation of the regression branch, i.e. the predictions for the slump flow δ as well as for the rheological parameters τ_0 and μ, the mean absolute error (MAE) and root mean squared error (RMSE) of the predictions are reported. In addition to the mean errors, we also compute the standard deviation of the absolute errors σ_{AE} to achieve further insights into the error distribution of the predictions.

Evaluation

This chapter provides the quantitative evaluation of the proposed Concrete Flow Transformer for the determination of fresh concrete properties based on open-channel flow observations. We report the results obtained by the *classification head* and the *regression head* separately.

Classification head

Fig. 5 shows the confusion matrix containing the results of the classification head predicting the consistency class of the concrete from the flow observations.

As is visible from the matrix, the classification leads to an overall accuracy (OA) of 77.4%, i.e. that many of the sliding window token sequences are associated the correct class by the Concrete Flow Transformer. The confusions show a clear pattern in the distribution of erroneous classifications, namely that the vast majority of errors appear next to the main diagonal, i.e. at neighbouring classes. A potential explanation for this observation is the fact that the consistency classes are obtained by a discretisation of the slump flow into individual consistency ranges, introducing class boundaries to the parameter range of the continuous variable δ. As several of concretes consistency classes lie very close to the class boundaries, potentially leading to an ambiguous setting of hard-to-distinguish concrete samples, partly causing the confusions that are observable in Fig. 5. While the classwise F1-scores range from 68.2% to 88.3% for four of the five classes, the class C_0 only achieves a F1-score of 54.8%. A reason for this can be the under-representation of this class in the data set leading to class-imbalance effects for the classification problem.

Consistency [mm]	soft 420-480	very soft 480-550	fluid 550-620	very fluid I 620-700	very fluid II >700	
Class	**C_0**	**C_1**	**C_2**	**C_3**	**C_4**	**Recall**
C_0	1.8%	0.2%	2.5%			39.6%
C_1		8.7%	3.2%			73.3%
C_2	0.2%	4.2%	32.7%	0.8%		86.2%
C_3			5.5%	11.5%	1.2%	63.3%
C_4			1.5%	3.2%	22.6%	82.7%
Precision	88.6%	66.6%	72.0%	74.0%	94.8%	
F1	54.8%	69.8%	78.5%	68.2%	88.3%	
OA	**77.4%**					

Figure 5: Confusion matrix for the classification head including the results for recall, precision, F1-score and overall accuracy. The values are to be read as percentage of entities belonging to reference class (rows) and being classified as the predicted class (columns). The colour coding denotes low values (light green) and high values (dark green). Empty entries correspond to values of 0.0%.

Regression head

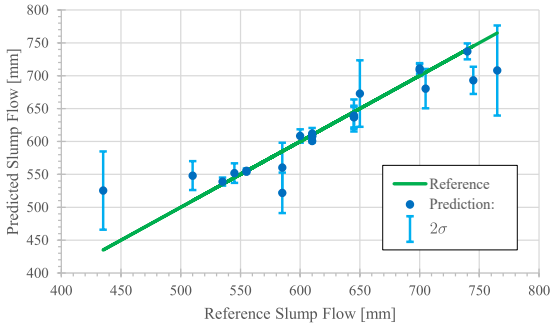
The regression head predicts the values for slump flow δ, the yield stress τ_0, as well as the viscosity μ. The quantitative results of the regression head predictions are shown in Tab. 1 The table contains the mean absolute error (MAE) of the target parameters, the standard deviation of the mean error σ_{AE} and the RMSE.

Table 1: Quantitative results obtained by the regression head. The MAE, the standard deviations of the absolute errors, as well as the RMSE are shown for the individual target parameters.

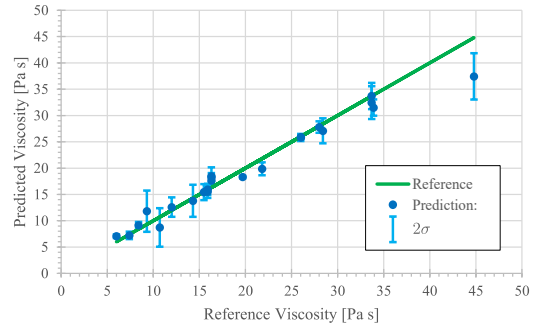
Parameter	MAE	σ_{AE}	RMSE
δ [mm]	23.3	34.5	41.6
τ_0 [Pa]	11.9	12.9	17.5
μ [Pa·s]	1.9	2.2	2.9

As is visible from the table, we achieve highly promising results with mean absolute errors of only 23.3 [mm], 11.9 [Pa], and 1.9 [Pa·s] for the respective target parameters, namely the slump flow diameter, yield stress, and plastic viscosity. As can be deducted from the relatively large standard deviations of the MAE, the errors exhibit a relatively broad distribution. In order to get deeper insights into the distribution of the predictions for the individual concrete samples, Fig. 6 shows the average predictions together with the corresponding standard deviation σ for each of the target parameters.

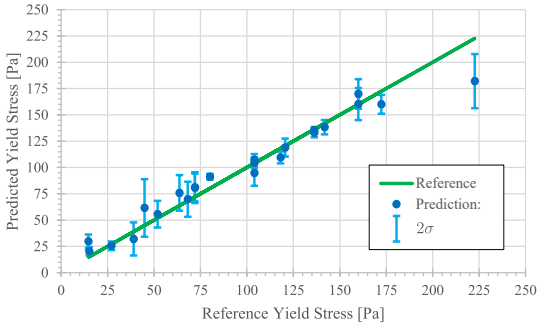
As is visible from the graphs of all three parameters, most of the mean predictions are very close to the reference diagonal, indicating that the prediction performance can be improved by averaging multiple predictions over time. An exception is the concrete sample with a plastic viscosity and yield stress at the maximum boundaries, and a slump flow at the minimum boundary of the value range, respectively. Here the distinctly largest difference between the mean prediction and the reference values and the largest



(a) Results for the slump flow δ.



(b) Results for the viscosity μ.



(c) Results for the yield stress τ_0.

Figure 6: Average predictions and the corresponding standard deviations σ of the target parameters δ, τ_0, and μ for each individual concrete sample contained in the test set.

standard deviation of the predictions are observable. We believe that this behaviour is caused by what is called the long-tail problem (Wang et al., 2017), which refers to the problem of an imbalanced statistical distribution of the examples in the training data where only little data is available for values in the tail of the distribution. This results in problems of these approaches w.r.t. their ability to generalise to barely or never seen examples or situations. This issue is currently being addressed by additional data acquisition series.

Conclusion

We presented the *Concrete Flow Transformer*, a novel method based on Vision Transformers for an automatic characterisation of fresh concrete from image sequences observing the open-channel concrete flow during the discharge process of a mixing vehicle. In this context, we proposed *Flow tokenisation* as efficient representation of patches serving as input to a transformer architecture. We showed that this representation encodes and preserves all information relevant for the derivation of the fresh concrete properties while drastically reducing the amount of (redundant) data. We demonstrated highly promising results by applying and evaluating the proposed method to a challenging real-world data set. In the future, we aim at building on research from the fields of fluid mechanics and flow modelling in order to incorporate knowledge from these fields to the problem of fresh concrete characterisation. In particular, we strive at incorporating sophisticated flow models as prior knowledge to our deep learning based approach, a strategy termed (physics) informed machine learning (von Rueden et al., 2021). We believe in this way to further improve the approach and to enhance the performance of the concrete characterisation.

Acknowledgements

The authors acknowledge the funding of the project *ReCyCONtrol* (`https://www.recycontrol.uni-hannover.de/en/`) provided by the German Federal Ministry of Education and Research (BMBF) under the grant No. 033R260A and the funding of the project *Open Channel Flow* provided by the German Research Foundation (DFG) under the grant No. 452024049.

References

Camgoz, N. C., Hadfield, S., Koller, O., and Bowden, R. (2016). Using Convolutional 3D Neural Networks for User-independent continuous Gesture Recognition. In International Conference on Pattern Recognition (ICPR), pages 49–54.

Coenen, M., Schack, T., Beyer, D., Heipke, C., and Haist, M. (2022). ConsInstancy: Learning Instance Representations for Semi-Supervised Panoptic Segmentation of Concrete Aggregate Particles. Machine Vision and Applications, 33(57).

DIN 1045-2 (2008). Concrete, Reinforced and Prestressed Concrete Structures - Part 2: Concrete - Specification, Properties, Production and Conformity - Application Rules for DIN EN 206-1.

Ding, Z. and An, X. (2018). Deep Learning Approach for Estimating Workability of Self-Compacting Concrete from Mixing Image Sequences. Advances in Materials Science and Engineering, 2018:1–16.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2625–2634.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.

(2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations (ICLR).

EN 12350-5 (2019). Testing Fresh Concrete - Part 5: Flow Table Test. European Committee for Standardization.

EN 206-1 (2001). Concrete - Part 1: Specification, Performance, Production and Conformity.

Farnebäck, G. (2003). Two-Frame Motion Estimation based on Polynomial Expansion. In Scandinavian Conference on Image Analysis, pages 363–370.

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016). Convolutional Two-Stream Network Fusion for Video Action Recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1933–1941.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. MIT Press, Cambridge, Massachusetts, USA.

Haist, M., Heipke, C., Beyer, D., Coenen, M., Vogel, C., Schack, T., Ponick, A., and Langer, A. (2022). Digitization of the Concrete Production Chain using Computer Vision and Artificial Intelligence. In Proceedings of the 6th fib Congress, pages 434–443.

Haist, M., Link, J., Nicia, D., Leinitz, S., and et al. (2020). Interlaboratory Study on rheological Properties of Cement Pastes and Reference Substances: Comparability of Measurements performed with different Rheometers and Measurement Geometries. Materials and Structures, 53(92).

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8):1735–1780.

Ji, S., Xu, W., Yang, M., and Yu, K. (2012). 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 35(1):221–231.

Kingma, D. and Ba, L. (2015). Adam: A Method for Stochastic Optimization. In International Conference on Learning Representations (ICLR).

Ponick, A., Langer, A., Beyer, D., Coenen, M., Haist, M., and Heipke, C. (2022). Image-Based Deep Learning for Rheology Determination of Bingham Fluids. In International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLIII-B2-2022, pages 711–720.

Sigurdsson, G. A., Divvala, S., Farhadi, A., and Gupta, A. (2017). Asynchronous Temporal Fields for Action Recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 585–594.

Simonyan, K. and Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. In Advances in Neural Information Processing Systems (NIPS), pages 568–576.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. In IEEE International Conference on Computer Vision (ICCV), pages 4489–4497.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In Advances in Neural Information Processing Systems (NIPS), volume 30.

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Walczak, M., Pfrommer, J., Pick, A., Ramamurthy, R., Garcke, J., Bauckhage, C., and Schuecker, J. (2021). Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. IEEE Transactions on Knowledge and Data Engineering.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. (2018). Temporal Segment Networks for Action Recognition in Videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 41(11):2740–2755.

Wang, X., Zhang, S., Qing, Z., Shao, Y., Zuo, Z., Gao, C., and Sang, N. (2021). OadTR: Online Action Detection with Transformers. In IEEE International Conference on Computer Vision (ICCV), pages 7565–7575.

Wang, Y.-X., Ramanan, D., and Hebert, M. (2017). Learning to Model the Tail. In Advances in Neural Information Processing Systems (NIPS), volume 30, pages 7029–7039.

Whipple, K. X. (1997). Open-Channel Flow of Bingham Fluids: Applications in Debris-Flow Research. The Journal of Geology, 105(2):243–262.

Yahia, A., Mantellato, S., and Flatt, R. J. (2016). Concrete Rheology: A Basis for Understanding Chemical Admixtures. In Science and Technology of Concrete Admixtures, pages 97–127. Woodhead Publishing.

Yang, H., Jiao, S.-J., and Yin, F.-D. (2020). Multilabel Image Classification Based Fresh Concrete Mix Proportion Monitoring Using Improved Convolutional Neural Network. Sensors, 20(16).

Zhou, L., Zhou, Y., Corso, J. J., Socher, R., and Xiong, C. (2018). End-to-End Dense Video Captioning with Masked Transformer. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 8739–8748.