Proceedings of Manufacturing Science and Engineering Conference
MSEC 2020
June 22-26, 2020, Cincinnati, Ohio, USA

MSEC-2020-13218

SPATIOTEMPORAL FUSION NETWORK FOR THE DROPLET BEHAVIOR RECOGNITION IN INKJET PRINTING

Jida Huang
Mechanical and Industrial Engineering
University of Illinois at Chicago
Chicago, Illinois 60607

Tianjiao Wang, Luis Javier Segura,
Gayatri Shirish Joshi, Hongyue Sun, Chi Zhou*
Industrial and Systems Engineering
University at Buffalo, SUNY
Buffalo, New York 14260
*Email: chizhou@buffalo.edu

ABSTRACT

Inkjet 3D printing has broad applications in areas such as health and energy due to its capability to precisely deposit micro-droplets of multi-functional materials. However, the droplet of the inkjet printing has different jetting behaviors including drop initiation, thinning, necking, pinching and flying, and they are vulnerable to disturbance from vibration, material inhomogeneity, etc. Such issues make it challenging to yield a consistent printing process and a defect-free final product with desired properties. Therefore, timely recognition of the droplet behavior is critical for inkjet printing quality assessment. In-situ video monitoring of the printing process paves a way for such recognition. In this paper, a novel feature identification framework is presented to recognize the spatiotemporal feature of in-situ monitoring videos for inkjet printing. Specifically, a spatiotemporal fusion network is used for droplet printing behavior classification. The categories are based on inkjet printability, which is related to both the static features (ligament, satellite, and meniscus) and dynamic features (ligament thinning, droplet pinch off, meniscus oscillation). For the recorded droplet jetting video data, two streams of networks, the frames sampled from video in spatial domain (associated with static features) and the optical flow in temporal domain (associated with dynamic features), are fused in different ways to recognize the droplet evolving behavior. Experiments results show that the proposed fusion network can recognize the droplet jetting behavior in the complex printing process and identify its printability with learned knowledge, which can ultimately enable the real-time inkjet printing quality control and further provide guidance to design optimal parameter settings for the inkjet printing process.
Keywords: Inkjet Printing, Spatiotemporal Fusion Network, Process Monitoring

1 Introduction

Inkjet printing is a direct depositing technique that is realized by ejecting liquid-phase materials (i.e., solutions/inks at different concentrations) to the substrate to form the final product. It has been extensively deployed in material patterning for the fabrication of functional parts, such as sensors, optic/electronic devices, biochips, among others [1], and has broad applications in health, energy, environment and electronics areas [2–4]. Among different inkjet printing processes, the Drop-On-Demand (DOD) method can achieve the highest resolution reported so far [5]. A suitable technology to supply droplets in DOD mode is the Piezoelectric Inkjet (PIJ) process (see Figure 1). In the PIJ process, the droplet formation is governed by tuning the driving electrical signal, various solution/ink properties (e.g., surface tension, viscosity, etc.), and the interaction between solution/ink, air, and substrate (e.g., wettability of the nozzle) [6, 7].

The droplet formation and ejection of PIJ will determine the properties of the final product. One major challenge of this process is that a droplet can have different jetting behaviors including drop initiation, thinning, necking, pinching and flying, and they are vulnerable to the variations such as vibration, material inhomogeneity, etc. [8]. The deposition rate in PIJ is typically

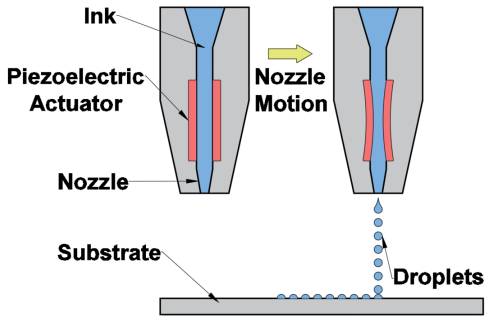


Figure 1: A schematic representation of the piezoelectric inkjet technology

100 to 1000 droplets per second, making it challenging to manually keep track of each individual droplet. Therefore, a real-time monitoring system would be very helpful to understand the droplet behavior under such high frequency deposition process and further identify the defects of printed parts. To build such a monitoring system, in the literature, many researchers have studied the droplet ejection and pinching behaviors by various approaches.

High-resolution and high-speed cameras are typically used to capture the droplet shape as it evolves over time [8, 9]. Such experimental characterization of the droplet behavior usually requires additional devices to estimate nozzle size, voltage signal, jetting speed, droplet shape, and ink properties [8]. For example, dimensionless parameters such as Reynolds (*Re*, ratio inertial-to-viscous forces), Weber (*We*, inertia-to-surface tension) are used to characterize the droplet evolution in [10]. Although the dimensionless numbers are helpful for droplet pinch-off estimation, they need a lot of experimental observations and the empirical results for one material and experimental setting can hardly be generalized to other materials and settings. In addition, the experimental approach is usually limited by a rough estimation based on visualizing the exterior of the inkjet nozzle, and ignores the complex processes that happen before the droplets are ejected.

Therefore, a real-time droplet behavior recognition system is needed. We hypothesize that an offline learning model, trained based on the limited amount of labelled experimental data, could be used for the online recognition of the high-speed droplet formation process. We use the state-of-the-art machine learning approach, Convolutional Neural Networks (ConvNet) [11], which has been extensively used in areas such as computer vision and image analysis with remarkable performance. In addition, as the collected data for droplet behavior is in video format, we are not only interested in learning the shape evolution of the droplet, its motion in temporal domain is also of interest. Therefore, a spatiotemporal fusion ConvNet which includes both spatial and temporal ConvNet is applied for the droplet video data training. The technical contributions of the paper are summarized as follows:

1. We investigated the intrinsic properties of the droplet forming process and proposed optical flow technique to extract the temporal information of the video data, and used the state-of-the-art ConvNet for the temporal feature learning.
2. We designed and implemented a spatiotemporal fusion framework for inkjet printing classification, which enables in-situ monitoring and process control towards a certified quality control system for additive manufacturing.

The remaining parts of the paper is organized as follows. Section 2 briefly review the related works. In Section 3, we discuss the PIJ droplet monitoring system for the droplet data collection as well as the optical flow calculation. The network architecture is presented in Section 4. Section 5 shows the experimental results from the designed fusion network. Finally, Section 6 concludes and summarizes the paper.

2 Literature Review

In this section, we will briefly discuss the related literature about the 3D printing process monitoring and machine learning with video data.

2.1 3D printing process monitoring

Monitoring and controlling the droplet formation process, which is a critical quality-determining factor in inkjet printing, are crucial to improve quality, repeatability, and consistency of the printed parts. Researchers have investigated new forms of instrumentation and adaptive approaches to further enhance the quality of the printed parts [12] Among various instruments, the image and video based devices such as cameras are most widely used. For instance, the droplet formation process of low, medium, and high viscosity inks was investigated in [13] by recording videos with a high-speed camera. Additionally, a new monitoring system that can show, within 2 *seconds*, the jetting status of a piezo driven inkjet printhead was proposed in [14]. A charged coupled device (CCD) camera was utilized to obtain the images along with the implementation of a low cost monitoring module that can measure the piezo self-sensing signals. Through the ink droplet images analyses, which were performed by deploying the edge detection technique, the jetting condition could be visualized and compared with the monitored results based on the piezo self-sensing signals. A closed-loop control framework by seamlessly integrating vision-based technique and neural network to inspect droplet behaviors and accordingly stabilize the printing process was presented in [9]. For the reviewed image- and video-based works, the complete droplet formation process is not fully considered. Furthermore, deep learning techniques, specifically a multi-scale convolutional neural network, for autonomous detection and classification of anomalies based on im-

ages has been developed in [15] for the Laser Powder Bed Fusion (LPBF). High-speed image acquisition, coupled with image segmentation and feature extraction, is used to estimate different statistical descriptors of the spattering behavior along the laser scanning path [16]. However, these methods failed to capture the complete process since they focus on specific stages of the LPBF process, and can hardly be incorporated to the full droplet formation in the PIJ process. Numerical simulations have been used to close this gap instead [6, 8, 17], although these methods are computational expensive and infeasible for real-time monitoring and control.

2.2 Video learning

Video has been studied for decades in the area of computer vision. Different types of problems such as action detection and recognition were studied with video data. For example, the action recognition based on the local image features or interest points were presented in [18]. Other tasks like anomaly detection [19] were explored by various researchers. Different shape descriptors in image domain such as scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG) were extended to 3D space [20, 21]. Dollar et al. proposed cuboids features for behavior recognition [22]. Wang et al. improved the performance of dense trajectories by taking into account camera motion to correct them [23]. Most of these works used hand-crafted features to extract the expected information from videos, hence they heavily relied on the pre-selected or defined features which usually is computationally intensive and becomes intractable on large-scale datasets.

With the breakthrough performance, ConvNets has been widely applied to different type of tasks in both images and videos [24, 25]. Le et al. proposed independent subspace analysis method for learning hierarchical invariant spatio-temporal features in action recognition [26]. While in [27], a novel 3D CNN model for action recognition was developed. The 3D ConvNet was also used to generate affinity graphs for medical image segmentation [28]. The Restricted Boltzmann Machines is combined with 3D CNN to learn spatio-temporal features [29]. Recently, Karpathy et al. utilized an extensive empirical evaluation of ConvNets on large-scale video classification using a large dataset [30]. Tran et al. proposed a model which performs 3D convolutions and 3D pooling propagating temporal information across all the layers in the network. However, such network is considerably deeper than the previous works [31]. Simonyan et al. investigated a two-stream architecture of discriminating trained network for action recognition in video [32]. Such a spatio-temporal network achieves superb performance for human action and is extensively exploited in different works [33]. In this paper, this method is introduced to the inkjet printing process through video learning. The fused network is used for the real-time monitoring of the inkjet printing process and identifying potential quality issues.

3 Data Collection System

In this section, the monitoring data collection system and optical flow is briefly introduced.

3.1 Monitoring data collection

Experiments are conducted to collect the video data for the training and testing of the network. The whole hardware setup is shown in Figure 2. In this system, a piezo-based micro-dispensing nozzle (MicroFab Inc.) is used as the inkjet droplets generation device, it has a nozzle size of 50 μm and can be operated with a jetting rate from 100 to 1000 droplets/second. The piezo-based micro-dispenser is driven by symmetrical trapezoid voltage. In the experiments, the peak drive voltage and back-pressure are selected as the design variables. Specifically, the voltage varies from 30 to 70V, and the back-pressure varies from -1 to -5 inch-water. Owing to their excellent rheological properties favorable for inkjet printing, typical Newtonian materials (deionized water and isopropyl alcohol in this study) are used in the experiments. A CCD camera (Sensor Technologies Inc.) coupled with a magnification lens works as the video capturing device, each frame obtained has a resolution of 640×480 pixels, and the data are transferred to the computer through USB protocol. To collect videos of the droplets generation process, strobing technology is utilized [34], which is also known as synchronized illumination technology and implemented by synchronizing the droplet jetting signal and the lighting signal of the LED for illumination. By tuning the delay time between these two signals, when capturing the repeated droplets generation process, the time between every two frames can be set precisely. In the data collection experiments, we set the delay time as 20 μs.

From the above data collection system, we can collect the vision data of the droplet forming process. A sample video is shown in Figure 3. Through analyzing such collected vision data, we can capture the evolution of the droplet and recognize its forming behavior, which can be further utilized for process monitoring and printability analysis.

3.2 Optical flow of droplet ejecting

To make use of the temporal information of the droplet forming in the video data, the optical flow is used as the input of the neural network in this work. The optical flow is defined as the distribution of apparent velocities of object movement pattern in an image [35]. It represents the movement of the observing object, specifically the dynamic behavior of the droplet in this paper. A dense optical flow can be seen as a set of displacement vector fields d_t between the pairs of consecutive frames t and $t+1$. We denote $d_t(u,v)$ as the displacement vector at the point (u,v) in frame t to the new point in the next frame $t+1$. The

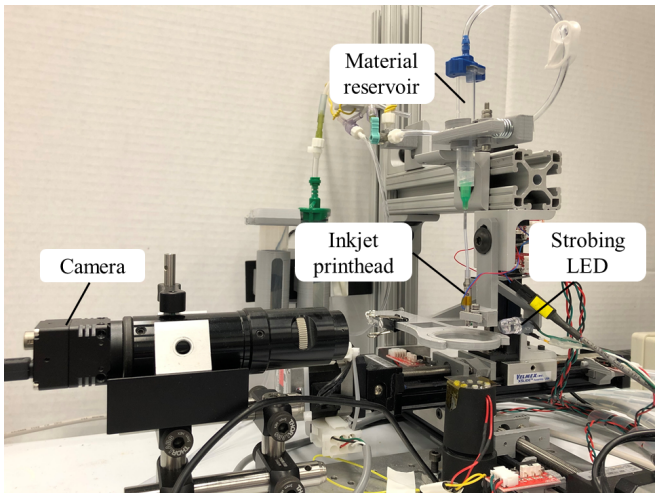


Figure 2: Inkjet printing device and data collection system

horizontal and vertical components of the vector field d_t^x and d_t^y can be seen as the image channels. An example of the optical flow in the droplet forming behavior in the PIJ process is shown in Figure 4.

The optical flow is computed through Horn-Schunck Method [35]. It can be seen from Figure 4 that the optical flow can capture the accelerated velocity and direction of each point in the image domain, which reflects the movement of the droplet. Figure 4(a) and (b) show two consecutive frames corresponding to drop initiation and pre-pinching stages. Based on these two video frames, Figure 4(c) shows the derived optical flow and visualizes the motion of the droplet. The slower motion in the left region (the cone area close to the orifice of the nozzle) indicates a concentric optical flow, while the faster motion in the right region (the neck of the droplet) shows centrifugal direction. Since the optical flow depicts the motion of the droplet, it can be used to characterize the dynamic behavior and accordingly determine the ejecting quality of the droplet. In the next section, we will discuss how to integrate the optical flow as temporal information into the network architecture.

4 Two-stream Fusion Network for Droplet Forming Behavior Recognition

Video can be naturally decomposed into spatial and temporal components. The spatial part, in the form of individual frame appearance, carries information of the scenes and objects depicted in the video. The temporal part, in the form of motion across the frames, conveys the movement of the observer (the camera) and the objects. Specifically, for the inkjet printing process studied in this work, the spatial component is related to the static features (ligament, satellite, and meniscus), and the temporal component is related to the dynamic features (ligament thinning, droplet pinch off, and meniscus oscillation). In this section, we will introduce the two-stream network architectures for the droplet video data learning, and the input and output data format of the training network.

4.1 Data labeling

In order to recognize the droplet behavior in the printing process, the intrinsic features embedded in the video data should be learned and extracted by the network. These features should be distinguishable from each other in order to avoid ambiguity during classification and recognition of different droplets. Hence, clearly labeled data should be provided for the network to learn such features. The aim of this paper is to identify the inkjet printability by recognizing the droplet forming behavior. It follows that the level of the printability can be used as the main factor for the data labeling. Based on the experimental data collected, we classify the dataset into four types as shown in Table 1.

Among the four different types of droplet forming behavior, the "excellent printability" is assigned to the droplets with clear pinching and very few tiny satellites followed, and the primary droplet contains the majority of the volume. For the ones with "good printability", the primary droplet is also pinched with satellites, but the secondary drops are connected with the tail satellite of the primary drop. In most cases, such behavior does not affect the overall printing quality, they are however vulnerable to external disturbance. The "fair printability" is for the droplets with unclear pinching and heavy broken satellites, as well as unstable flying trajectory. The drops with "fair printability" will severely affect the quality of the printed part and should be corrected or abandoned in real applications. The last case is classified as "poor printability", mainly for the droplets that can not be properly formed, which are typically caused by low back-pressure, low driving voltage, high material viscosity and nozzle clogging etc. Based on such classification criterion, all of the collected video data are labeled and provided for the network to learn the features of the jetting behavior.

4.2 Spatio-temporal fusion network for video recognition

Most of the existing neural network frameworks use the images as input, and the success of the convolutional neural network benefits from the hierarchical feature learning ability in spatial domain. However, for the video data, the motion feature in temporal domain is also an important factor for the object recognition. Hence, the temporal feature should be considered especially for the tasks relying on the motion information.

In this paper, in order to learn the droplet evolving behavior, a two-stream (in both spatial and temporal domain) neural network is built for the feature understanding of the PIJ printing process. The input data of the network are formatted with the dimension of $H \times W \times C \times L$, where L is the number of frames, H and W are the height and width of each frame, C is the number

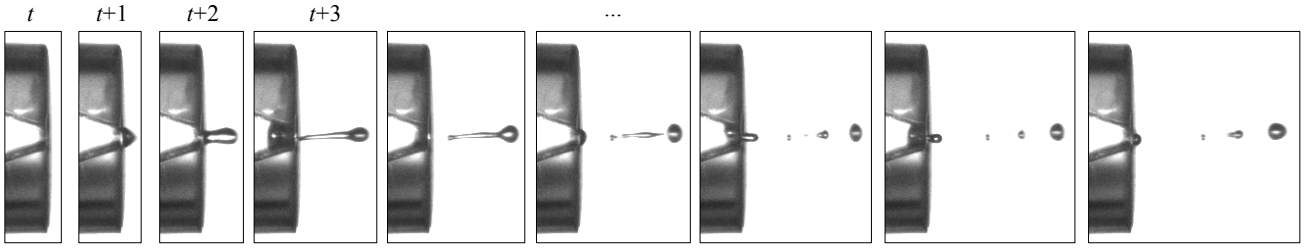


Figure 3: Multiple consecutive time frames $(t, t+1, t+2, ...)$ from the collected sample video data.

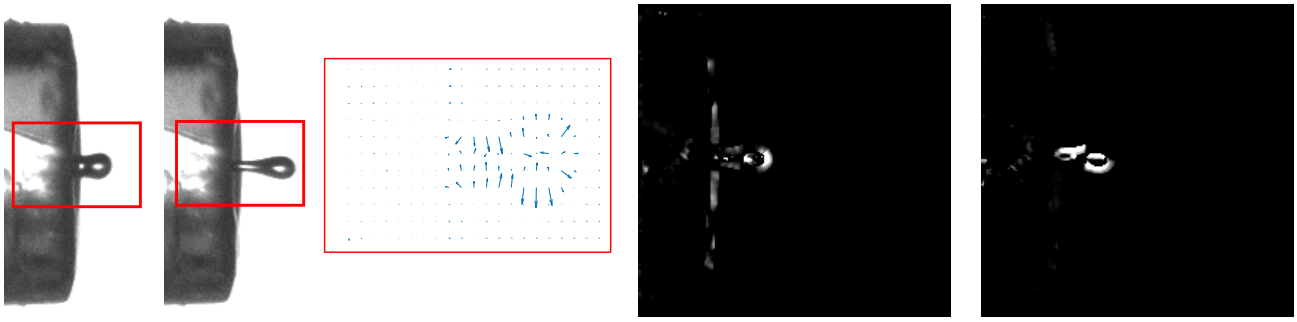


Figure 4: Optical flow. (a)-(b): a pair of consecutive video frames with the droplet ejecting outlined with a red frame. (c): a close-up view of the dense optical flow in the outlined area; (d)-(e): horizontal component d^x and vertical component d^y of the displacement vector field, higher intensity corresponds to positive value and lower intensity corresponds to negative value.

of channels of the frames. The output data of the network are the category labels of the video. In this work, the dimension of the video clips is $216 \times 216 \times 1 \times 10$, i.e. 10 consecutive frames are sampled from the droplet video, each frame is cropped and down-sampled to 216×216 pixels. Here the video is recorded in gray-scale mode, hence the number of channels is 1. As discussed before, the droplet forming behaviors are classified into four different types, hence the output data have four different class labels. The two different streams of the networks are introduced in the following sections.

4.2.1 Spatial stream ConvNet

Spatial stream ConvNet operates on individual droplet jetting video frames, which performs droplet behavior recognition in the printing process from these static images. The static appearance of the droplet in the jetting process provides a useful clue, since many type of droplets visually form into different shapes, and such shapes are strongly associated with a particular type of printability.

Since the spatial ConvNet is essentially an image classification architecture, we can make use of the recent advances in large-scale image recognition methods. In this paper, we applied a state-of-the-art pre-trained deep network, ResNet50 [36], with fine-tuning in our spatial network architecture for the droplet behavior recognition. The architecture is depicted in Figure 5. 10 consecutive frames (i.e., a $216 \times 216 \times 1 \times 10$ volume) are sampled from each video for the input of the spatial stream ConvNet. Three fully connected layers with dimensions of 4096, 2048 and 4 are concatenated to the ResNet50 to achieve droplet behavior classification. The softmax layer is used on the classification layer.

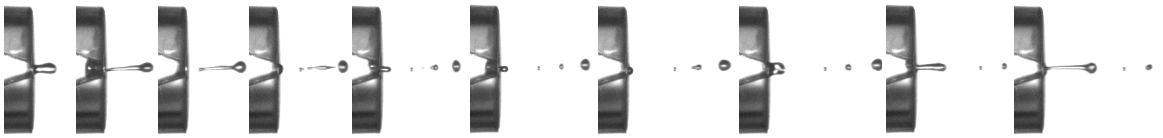
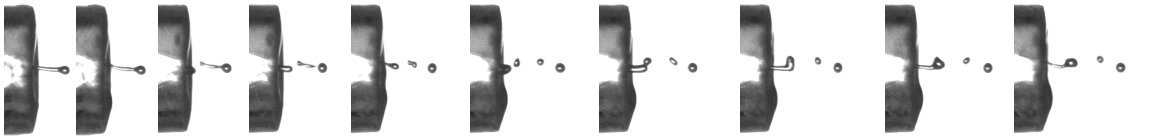
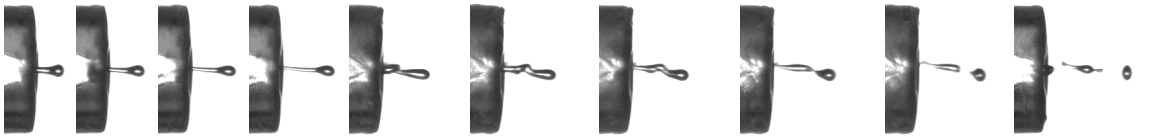
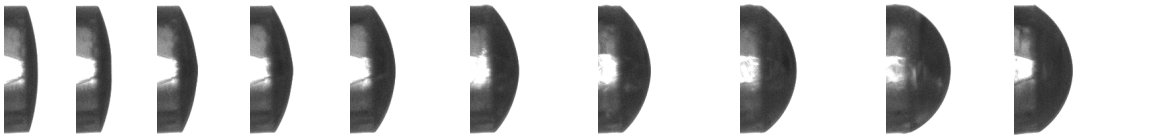
4.2.2 Temporal stream ConvNet

Optical flow stacking. To represent the motion across a sequence of frames, we stack the flow channels $d_t^{x,y}$ of L consecutive frames to form a total of $2L$ input channels, here $d_t^{x,y}$ is the displacement vector at the point (x,y) in frame t, which moves the point to the corresponding point in the following frame t + 1. More formally, let H and W be the height and width of a video, a ConvNet input volume $I_\tau \in \mathbb{R}^{H \times W \times 2L}$ for an arbitrary frame τ is then constructed as follows:

$$\begin{aligned} I_\tau(i,j,2k-1) &= d^x_{\tau+k-1}(i,j) \\ I_\tau(i,j,2k) &= d^y_{\tau+k-1}(i,j) \\ 1 \le i \le H, 1 \le j \le W, 1 \le k \le L \end{aligned} \tag{1}$$

For an arbitrary point (i,j), the channels $I_\tau(i,j,c)$; $c = [1 : 2L]$ encode the motion at that point over a sequence of L frames.

Table 1: Four different type of droplet jetting behaviors

Label	Example
Excellent printability	
Good printability	
Fair printability	
Poor printability	

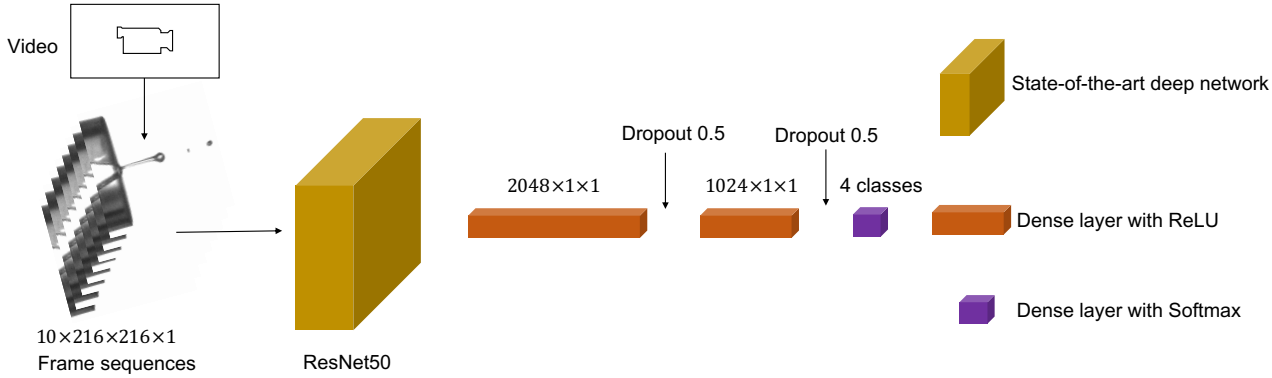


Figure 5: Spatial stream ConvNet architecture.

After stacking multiple optical flow displacement fields into a single volume $I_\tau \in \mathbb{R}^{H\times W\times 2L}$, we feed it into the ConvNet. Considering that a ConvNet requires a fixed-size input, we sample 10 consecutive frames from the video and compute 9 optical flow between these frames. Then a sub-volume of $216 \times 216 \times 18$ is formed and passed to the network as input. The hidden layers configuration remains similar as those used in the spatial net. The temporal stream ConvNet is illustrated in Figure 6.

The network has 5 convolution layers, 2 pooling layers, 3 fully-connected layers and a softmax loss layer to predict the jetting behavior. The first convolution layers are followed by a pooling layer. The number of filters for the 5 convolution layers from 1 to 5 are 96, 256, 512, 512, 512, respectively. All of these convolution layers are applied with appropriate padding (both spatial and temporal) and stride of 1, thus there is no change in term of size from the input to the output of all convolution layers. All pooling layers are max pooling with kernel size of $2 \times 2 \times 2$ with stride of 1. Three fully connected layers have 4096, 2048 and 4 outputs. Same as the spatial ConvNet, the softmax layer is used on the classification layer.

4.2.3 Two-stream fusion network architecture

Sections 4.2.1 and 4.2.2 introduced the spatial and temporal stream CovnNet for the droplet behavior recognition. These two networks only consider one-fold information of the inkjet printing process. In this section, we introduce a fusion network that combines the two aforementioned networks together to perform the droplet printability classification. The two-stream network

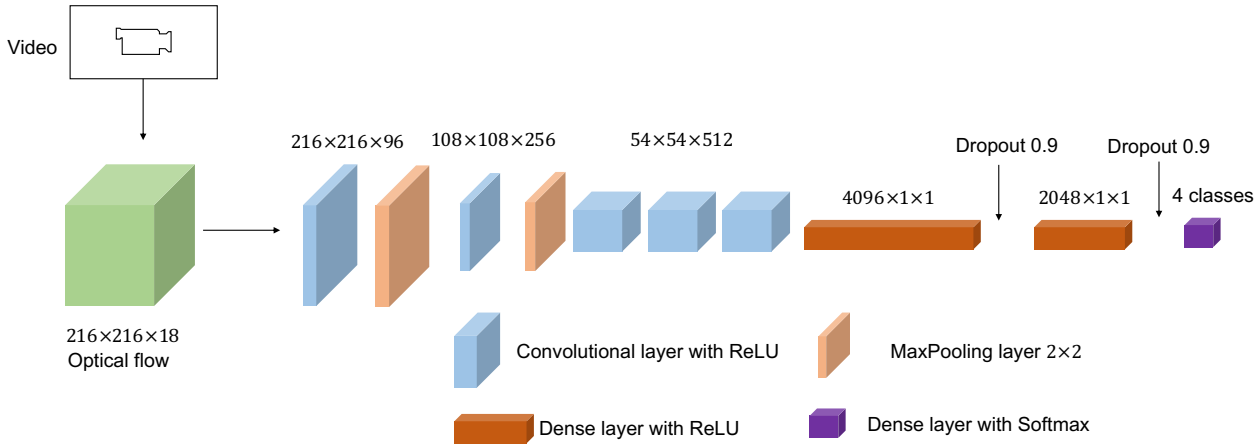


Figure 6: Temporal stream ConvNet by taking multi-frame optical flow as input.

architecture is shown in Figure 7.

Fusion method. As can be seen in Figure 7, the fusion function is performed in the last dense fully connected layers. It should be noted that the fusion can be applied at any point in the two networks. Suppose a fusion function f, $y_t = f(x_t^a, x_t^b)$ fuses two feature maps $x_t^a \in \mathbb{R}^{H1\times W1\times C1}$ and $x_t^b \in \mathbb{R}^{H2\times W2\times C2}$, at time t, to produce an output map $y_t \in \mathbb{R}^{H\times W\times C}$, where H, W and C are the height, width and number of channels of the respective feature maps. For the sake of simplicity, we assume that $H = H1 = H2, W = W1 = W2, C = C1 = C2$. When applied to ConvNet architectures, consisting of convolutional, fully connected, pooling and non-linearity layers, f can be applied at different locations in the network, e.g., early-fusion, late-fusion or multiple layer fusion [33]. Various fusion functions f can be used. In this paper, we mainly consider two type of fusion methods.

Average fusion. Average fusion computes the average of the two feature maps at the same spatial locations i, j and feature channels d.

$$y_{i,j,d} = \frac{x^a_{i,j,c} + x^b_{i,j,c}}{2} \tag{2}$$

where $1 \le i \le H, 1 \le j \le W, 1 \le c \le C$, and $x^a, x^b, y \in \mathbb{R}^{H\times W\times C}$.

Conv fusion. Conv fusion first stacks the two feature maps at the same spatial locations i, j across the feature channels c and subsequently conducts a convolution operation on the stacked data with a bank of filters $\mathbf{f} \in \mathbb{R}^{1\times 1\times 2C\times C}$ and biases $b \in \mathbb{R}^C$.

$$y = Concatenation(x^a, x^b) * \mathbf{f} + b \tag{3}$$

where the number of output channels is C. When used as a trainable filter kernel in the network, $\mathbf{f}$ is able to learn the correspondence of the two feature maps that minimize a joint loss function.

5 Experimental Results and Discussion

In this section, the performance of the proposed spatiotemporal fusion network on the collected dataset is tested. We collect the monitoring data through the system introduced in Section 3.1. The fixed system parameters are set as in Section 3.1, then the experiments are conducted by changing the variable parameters including peak drive voltage and back-pressure, these variables are randomly selected with random combinations. In total we collect 4K videos of the process monitoring data. In the training stage, 3K videos are used for training the network, and 500 are used for validation and testing, respectively. The layer configuration of the spatial and temporal ConvNets are detailed in Figures 5-7. All hidden weight layers use the rectification (ReLU) activation function, maxpooling is performed over 2×2 spatial windows with stride 1.

In this experiment, we test the performance of different types of networks described in Section 4 on various sizes of training data. The training procedure is conducted on the collected video frames, and is generally the same for both spatial and temporal nets. The network weights are learnt using the mini-batch stochastic gradient descent with momentum. At each iteration, a mini-batch of 200 samples is constructed by sampling 200 training videos (uniformly across the classes), from each of which a single frame is randomly selected. In spatial net training, a 216×216 sub-image containing the area of interest (i.e., the printing head and the droplet trajectory) is cropped from the selected frame. In the temporal net training, we compute an optical flow volume I for the selected training frame as described in Section 4.2.2. From that volume, a fixed-size $216 \times 216 \times 18$ input is stacked. The learning rate is initially set to 10^{-2} and then decreased according to a fixed schedule, which is kept the same for all training sets. Namely, when training a ConvNet from scratch, the rate is changed to 10^{-3} after 30K iterations, then to 10^{-4} after 40K iterations, and the training is terminated after 50K iterations. In the fine-tuning scenario, the rate is changed to

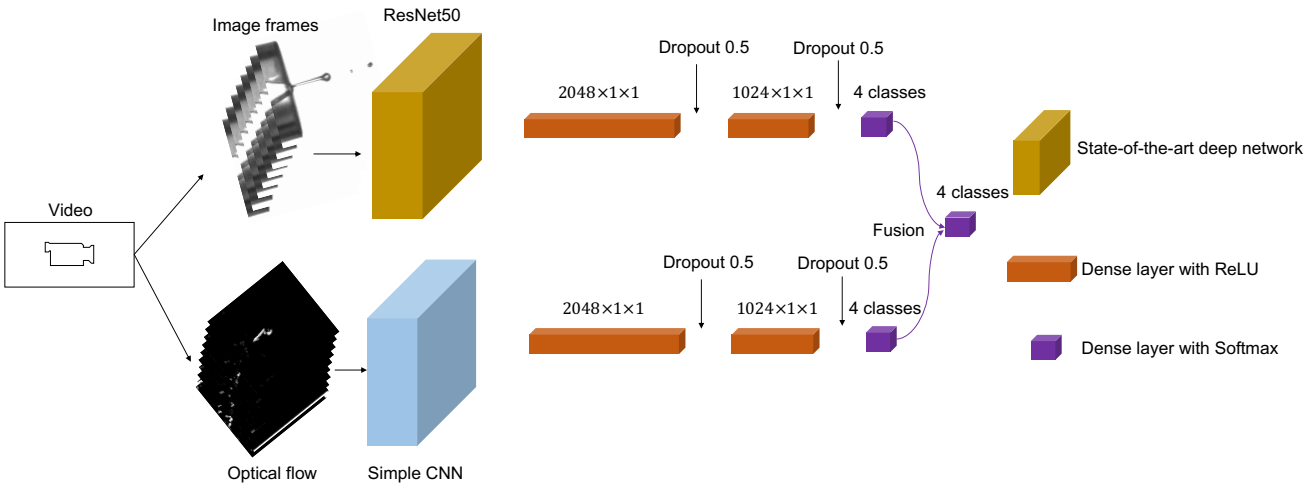


Figure 7: Two-stream network architecture for droplet behavior recognition.

Table 2: Classification results on the collected video data

Training data size	ResNet50	Temporal stream ConvNet	Fused ConvNet (Average fuse)	Fused ConvNet (Conv fuse)
3000	78.2%	70.0%	88.3%	**94.2%**
2000	76.3%	67.6%	84.5%	**93.5%**
1000	75.8%	66.0%	82.9%	**90.5%**
500	73.4%	62.6%	80.4%	**87.8%**

10^{-3} after 15K iterations, and training stopped after 20K iterations. The training is optimized using the Adam algorithm with $\beta_1 = 0.9, \beta_2 = 0.999$ [37]. Table 2 listed the droplet comparison results for behavior recognition accuracy on a set of same testing samples with different size of training data. The accuracy is measured as the correct printability prediction percentage of the input test data.

It can be seen form Table 2 that the performance of trained network is stable on different sizes of training data (i.e., the difference of the prediction results within training size is small). This reveals that these networks are stable for the droplet behavior recognition on the video data. It can also be observed that with the spatial stream (ResNet50) alone, the network can obtain a fair recognition accuracy. This is mainly because the forming shape of droplet pinching process is related to the final printing quality. Thus the ResNet50 can capture such shape feature to enable its droplet behavior recognition performance.

With the spatiotemporal fusion, we can see the network obtained a remarkably improved recognition accuracy. The accuracy of two-stream fused network achieves up to 17.2% better than the spatial network. This reveals that through fusing the temporal information into the network, the network could have a discriminating capability for the droplet evolving feature, thus increasing the droplet behavior recognition accuracy. In addition, the convolution fusion performances better than the simple average fusion, this suggests an early convolution fusion of the spatial feature and temporal feature could be aggregated and beneficial for the classification of the droplet forming printability. It should be noted that the average fusion method still outperforms the solely spatial-stream or temporal-stream method.

6 Conclusions

In this paper, a two-stream network is proposed for the recognition of the droplet jetting behavior in the PIJ printing process. The collected data is firstly labeled based on the printability, then based on the spatial as well as temporal stream of video data, a fused network is proposed for the jetting behavior recognition. Though a trained network, the droplet evolving behavior is recognized for a new video data, and experiments results show the proposed method can capture the droplet evolving feature and identify the printability of the droplet during printing. By using such a video recognition framework, the proposed work can be extended to the printing process monitoring using a vision-based system, which can be applied to other types of nozzles and materials, also the learned network can be further used in the quality

control of the inkjet printing process. Future work includes the learned feature visualization, analysis and application for video clustering and integrating the process control parameters in the network.

ACKNOWLEDGMENT

We acknowledge the support from the National Science Foundation (NSF) through CMMI-1846863 and the seed fund support from SMART. We thank the support provided by the Center for Computational Research at the University at Buffalo.

REFERENCES

[1] Sun, J., Bao, B., He, M., Zhou, H., and Song, Y., 2015. "Recent advances in controlling the depositing morphologies of inkjet droplets". *ACS applied materials & interfaces,* ***7***(51), pp. 28086–28099.

[2] Mironov, V., Boland, T., Trusk, T., Forgacs, G., and Markwald, R. R., 2003. "Organ printing: computer-aided jet-based 3d tissue engineering". *TRENDS in Biotechnology,* ***21***(4), pp. 157–161.

[3] Sirringhaus, H., Kawase, T., Friend, R., Shimoda, T., Inbasekaran, M., Wu, W., and Woo, E., 2000. "High-resolution inkjet printing of all-polymer transistor circuits". *Science,* ***290***(5499), pp. 2123–2126.

[4] Yan, P., Brown, E., Su, Q., Li, J., Wang, J., Xu, C., Zhou, C., and Lin, D., 2017. "3d printing hierarchical silver nanowire aerogel with highly compressive resilience and tensile elongation through tunable poisson's ratio". *Small,* ***13***(38), p. 1701756.

[5] Hoath, S. D., 2016. *Fundamentals of inkjet printing: the science of inkjet and droplets*. John Wiley & Sons.

[6] Wijshoff, H., 2010. "The dynamics of the piezo inkjet printhead operation". *Physics reports,* ***491***(4-5), pp. 77–177.

[7] Basaran, O. A., Gao, H., and Bhat, P. P., 2013. "Non-standard inkjets". *Annual Review of Fluid Mechanics,* ***45***, pp. 85–113.

[8] He, B., Yang, S., Qin, Z., Wen, B., and Zhang, C., 2017. "The roles of wettability and surface tension in droplet formation during inkjet printing". *Scientific reports,* ***7***(1), p. 11841.

[9] Wang, T., Kwok, T.-H., Zhou, C., and Vader, S., 2018. "In-situ droplet inspection and closed-loop control system using machine learning for liquid metal jet printing". *Journal of manufacturing systems,* ***47***, pp. 83–92.

[10] Xu, C., Zhang, Z., Fu, J., and Huang, Y., 2017. "Study of pinch-off locations during drop-on-demand inkjet printing of viscoelastic alginate solutions". *Langmuir,* ***33***(20), pp. 5037–5045.

[11] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P., 2017. "Geometric deep learning: going beyond euclidean data". *IEEE Signal Processing Magazine,* ***34***(4), pp. 18–42.

[12] Everton, S. K., Hirsch, M., Stravroulakis, P., Leach, R. K., and Clare, A. T., 2016. "Review of in-situ process monitoring and in-situ metrology for metal additive manufacturing". *Materials & Design,* ***95***, pp. 431–445.

[13] Yang, H., He, Y., Tuck, C., Wildman, R., and Hague, R., 2013. "High viscosity jetting system for 3d reactive inkjet printing". In Twenty Forth Annual International Solid Freeform Fabrication Symposium—An Additive Manufacturing Conference, pp. 505–513.

[14] Kwon, K.-S., Choi, Y.-S., Lee, D.-Y., Kim, J.-S., and Kim, D.-S., 2012. "Low-cost and high speed monitoring system for a multi-nozzle piezo inkjet head". *Sensors and Actuators A: Physical,* ***180***, pp. 154–165.

[15] Scime, L., and Beuth, J., 2018. "A multi-scale convolutional neural network for autonomous anomaly detection and classification in a laser powder bed fusion additive manufacturing process". *Additive Manufacturing,* ***24***, pp. 273–286.

[16] Repossini, G., Laguzza, V., Grasso, M., and Colosimo, B. M., 2017. "On the use of spatter signature for in-situ monitoring of laser powder bed fusion". *Additive Manufacturing,* ***16***, pp. 35–48.

[17] Tan, H., Torniainen, E., Markel, D. P., and Browning, R. N., 2015. "Numerical simulation of droplet ejection of thermal inkjet printheads". *International Journal for Numerical Methods in Fluids,* ***77***(9), pp. 544–570.

[18] Laptev, I., 2005. "On space-time interest points". *International journal of computer vision,* ***64***(2-3), pp. 107–123.

[19] Boiman, O., and Irani, M., 2007. "Detecting irregularities in images and in video". *International journal of computer vision,* ***74***(1), pp. 17–31.

[20] Scovanner, P., Ali, S., and Shah, M., 2007. "A 3-dimensional sift descriptor and its application to action recognition". In Proceedings of the 15th ACM international conference on Multimedia, ACM, pp. 357–360.

[21] Klaser, A., Marszałek, M., and Schmid, C., 2008. "A spatio-temporal descriptor based on 3d-gradients". In BMVC2008.

[22] Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S., 2005. "Behavior recognition via sparse spatio-temporal features". VS-PETS Beijing, China.

[23] Wang, H., and Schmid, C., 2013. "Action recognition with improved trajectories". In Proceedings of the IEEE international conference on computer vision, pp. 3551–3558.

[24] Jain, A., Tompson, J., Andriluka, M., Taylor, G. W., and Bregler, C., 2013. "Learning human pose estimation features with convolutional networks". *arXiv preprint arXiv:1312.7302*.

[25] Jain, A., Tompson, J., LeCun, Y., and Bregler, C., 2014. "Modeep: A deep learning framework using motion fea-

tures for human pose estimation". In Asian conference on computer vision, Springer, pp. 302–315.
[26] Le, Q. V., Zou, W., Yeung, S., and Ng, A. Y., 2011. "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis".
[27] 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,* ***35***(1), pp. 221–231.
[28] Turaga, S. C., Murray, J. F., Jain, V., Roth, F., Helmstaedter, M., Briggman, K., Denk, W., and Seung, H. S., 2010. "Convolutional networks can learn to generate affinity graphs for image segmentation". *Neural computation,* ***22***(2), pp. 511–538.
[29] Taylor, G. W., Fergus, R., LeCun, Y., and Bregler, C., 2010. "Convolutional learning of spatio-temporal features". In European conference on computer vision, Springer, pp. 140–153.
[30] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L., 2014. "Large-scale video classification with convolutional neural networks". In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732.
[31] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M., 2015. "Learning spatiotemporal features with 3d convolutional networks". In Proceedings of the IEEE international conference on computer vision, pp. 4489–4497.
[32] Simonyan, K., and Zisserman, A., 2014. "Two-stream convolutional networks for action recognition in videos". In Advances in neural information processing systems, pp. 568–576.
[33] Feichtenhofer, C., Pinz, A., and Zisserman, A., 2016. "Convolutional two-stream network fusion for video action recognition". In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1933–1941.
[34] Wang, T., Kwok, T.-H., and Zhou, C., 2017. "In-situ droplet inspection and control system for liquid metal jet 3d printing process". *Procedia Manufacturing,* ***10***, pp. 968–981.
[35] Horn, B. K., and Schunck, B. G., 1981. "Determining optical flow". *Artificial intelligence,* ***17***(1-3), pp. 185–203.
[36] He, K., Zhang, X., Ren, S., and Sun, J., 2016. "Deep residual learning for image recognition". In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
[37] Kingma, D. P., and Ba, J., 2014. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980*.