# Cell Event Detection in Phase-Contrast Microscopy Sequences From Few Annotations

Melih Kandemir[1], Christian Wojek[2], Fred A. Hamprecht[1]

1 Heidelberg University HCI/IWR, Germany
2 Carl Zeiss AG, Oberkochen, Germany

**Abstract.** We study detecting cell events in phase-contrast microscopy sequences from few annotations. We first detect event candidates from the intensity difference of consecutive frames, and then train an unsupervised novelty detector on these candidates. The novelty detector assigns each candidate a degree of surprise. We annotate a tiny number of candidates chosen according to the novelty detector's output, and finally train a sparse Gaussian process (GP) classifier. We show that the steepest learning curve is achieved when a collaborative multi-output Gaussian process is used as novelty detector, and its predictive mean and variance are used together to measure the degree of surprise. Following this scheme, we closely approximate the fully-supervised event detection accuracy by annotating only 3% of all candidates. The novelty detector based annotation used here clearly outperforms the studied active learning based approaches.

## 1 Introduction

Modern cell biology thrives on time lapse imaging where the fate of cells can be studied over time and in response to various stimuli. Phase-contrast microscopy is an important experimental technique in this area because it is non-invasive and allows the detection of events such as mitosis (cell division) or apoptosis (cell death). A fundamental challenge in phase-contrast microscopy images is the hardness of segmenting the cell boundaries accurately. Even though there have been attempts for segmenting cells in this type of images [5], the accuracy and the tolerance to imaging artifacts provided by the state-of-the-art are significantly below the reliability level. The suggested methods for automated analysis of phase-contrast sequences by-pass the segmentation step and jump directly to detecting events of interest from a heuristically generated set of candidate regions. Huh et al. [3] extract a candidate mitotic region from each large enough bright spot in a frame. Each candidate is represented by a Histogram of Oriented Gradients (HoG) feature set, passed through a binary classifier, and the classifier decisions are smoothed by a Conditional Random Field (CRF). In [4], apoptotic events of stem cells are detected by extracting candidate regions exploiting the image acquisition principles of phase-contrast microscopy, a rule-based filtering using application-specific heuristics, and a support vector classifier.

All the above methods require a fully-supervised training sequence. A time-lapse sequence consists of hundreds of events, which often occur simultaneously, making their annotation a tedious task. Hence, a way of reducing the annotation labor would save the biologist's precious time. This problem has first been investigated by Kandemir

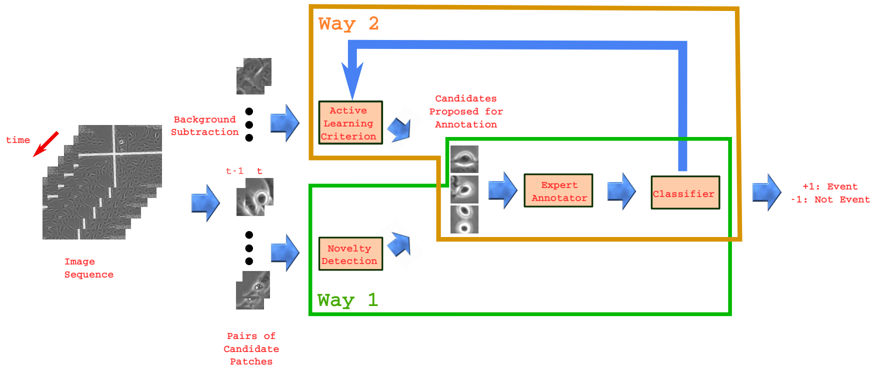


**Fig. 1.** Two proposed pipelines for interactive cell event detection. **Way 1:** Unsupervised event detection followed by supervised learning. **Way 2:** Active learning. The novelty detector we introduce in Way 1 gives a steeper learning curve than Way 2.

et al. [6], where a multioutput GP (MOGP) is used as an autoregressor to predict the features of a patch at time $t$ from its features at time $t - 1$. This autoregressor is trained on the first few (five) frames of the sequence where there occur no events of interest. The prediction error made by the autoregressor is used as the degree of surprise, and an event is fired if this surprise is above a preset threshold.

In this paper, we study ways of reaching high cell event detection accuracy using a supervised classifier as opposed to [6], but using only a small amount of annotations as opposed to [2,3,4]. For this, we introduce a new pipeline, and a new candidate generation scheme, which does not use labels as [2]. We then extract a set of generic features from each of these candidates, and train a novelty detector on the resulting data set. We assign a probability to each candidate proportional to its degree of surprise, and choose the candidates to be annotated by sampling from the resultant probability distribution. The annotated candidates are then fed into a binary classifier. The proposed pipeline is illustrated in Figure 1 as Way 1.

We test our pipeline and the proposed feature unpredictability based novelty detector on eight separately acquired sequences of human liver cell cultures, and three sequences of stem cell cultures. We have several interesting outcomes:

- Cell events can be detected almost as accurately as the fully-supervised case, and much more accurately than [6], by annotating less than 3% of the training data following the order suggested by the novelty detectors.
- The steepest learning curve is reached when a state-of-the-art collaborative multi-output GP regressor is used as the novelty detector, which is trained for predicting the feature vector of a candidate at frame $t$ from its feature vector at the previous frame $t - 1$, and the prediction error is calculated using the predictive mean and variance together, differently from [6] which uses only the predictive mean.
- Even though active learning (AL) looks like an intuitive solution to this problem (Figure 1 , Way 2), annotating the sequences using our unsupervised novelty detection (Way 1) performs clearly better than the studied algorithms.

## 2 Background Subtraction

As the common property of any event is an abrupt change in time, we propose taking the intensity difference of every consecutive frame, and fire a candidate event for each large enough connected component in the difference image. We threshold the intensity difference image from 25% of its brightest pixel value, filter out the components of the resultant binary image having an area smaller than 20 pixels, apply image opening with a 2-pixel disk, and finally fill the holes. Each connected component in the resultant image is treated as an event candidate. We represent a candidate by a 26-bin intensity histogram calculated on the intensity difference of that frame and its predecessor, 58 LBP features of the current frame, another 58 Local Binary Pattern (LBP) features for the difference from the predecessor, and 128 Scale Invariant Feature Transform (SIFT) features.

## 3 Sparse Gaussian Processes

Let $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ be a data set containing $N$ instances $\mathbf{x}_n \in \mathbb{R}^D$, and $\mathbf{y} = \{y_1, \cdots, y_N\}$ be the corresponding real-valued outputs. The main drawback of the standard GP is the inversion of the $N \times N$ kernel matrix, which undermines the scalability of the standard GP. There exist several sparse approximations in the GP literature to overcome this problem. Here, we choose the FITC approximation [13]

$$\begin{aligned} p(\mathbf{u}|\mathbf{Z}, \boldsymbol{\theta}) &= \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{ZZ}), \\ p(\mathbf{f}|\mathbf{X}, \mathbf{u}, \mathbf{Z}, \boldsymbol{\theta}) &= \mathcal{N}(\mathbf{f}|\mathbf{K}_{ZX}^T \mathbf{K}_{ZZ}^{-1} \mathbf{u}, diag(\mathbf{K}_{XX} - \mathbf{K}_{ZX}^T \mathbf{K}_{ZZ}^{-1} \mathbf{K}_{ZX})), \\ p(\mathbf{y}|\mathbf{f}) &= \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I}), \end{aligned} \qquad (1)$$

where $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, is the vector of noise-free latent outputs, $\mathbf{Z}$ is the pseudo input data set, $\mathbf{u}$ is the vector of its pseudo targets, $\mathbf{K}_{ZX}$ is the kernel responses between $\mathbf{Z}$ and $\mathbf{X}$, the vector $\boldsymbol{\theta}$ has the kernel hyperparameters, and $\sigma^2$ is the variance of the white noise on the outputs. The sparse GP, denoted as $\mathcal{SGP}(\mathbf{f}, \mathbf{u}|\mathbf{Z}, \mathbf{X})$, is a parametric approximation to the nonparametric full GP. For regression, $\mathbf{u}$ is available in closed form as $p(\mathbf{u}|\mathbf{Z}, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{u}|\mathbf{K}_{ZZ}\mathbf{Q}^{-1}\mathbf{K}_{ZX}(\boldsymbol{\Lambda} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \mathbf{K}_{ZZ}\mathbf{Q}^{-1}\mathbf{K}_{ZZ})$ where $\mathbf{Q} = \mathbf{K}_{ZZ} + \mathbf{K}_{ZX}(\boldsymbol{\Lambda} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{ZX}^T$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$ with $\lambda_n = k_{xx} - \mathbf{k}_{Zx}^T \mathbf{K}_{ZZ}^{-1} \mathbf{k}_{Zx}$. Here, $\mathbf{k}_{Zx}$ is the kernel responses between $\mathbf{Z}$ and a single instance $\mathbf{x}$. The marginal likelihood is $p(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{ZX}^T \mathbf{K}_{ZZ}^{-1} \mathbf{K}_{ZX} + \boldsymbol{\Lambda} + \sigma^2\mathbf{I})$. The remaining parameters $\mathbf{Z}$ and $\boldsymbol{\theta}$ are learned by gradient descent.

For binary classification, it suffices to replace the likelihood in Equation 1 with $p(\mathbf{t}|\mathbf{f}) = \prod_{n=1}^N \Phi(f_n)^{t_n}(1 - \Phi(f_n))^{1-t_n}$, where $\Phi(s) = \dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^s e^{-\frac{1}{2}s^2} ds$ is the probit link function and $\mathbf{t}$ is the vector of output classes $t_n \in \{-1, +1\}$. Since this likelihood is not conjugate with the normal distribution, the posterior $p(\mathbf{u}|\mathbf{Z}, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta})$ is no longer available in closed form, hence has to be approximated. We choose the Laplace approximation due to its computational efficiency and yet reasonable performance [9].

Given the approximate posterior $q(\mathbf{u}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})$, and a newly seen instance $(\mathbf{x}_n, t_n)$, the predictive distribution is $p(t_n|\mathbf{t}) = \int\int p(t_n|f_n) p(f_n|\mathbf{X}, \mathbf{u}, \mathbf{Z}) q(\mathbf{u}|\mathbf{X}, \mathbf{Z}, \mathbf{t}) d\mathbf{u} df_n$.

We follow the common practice and take only the first integral with respect to $\mathbf{u}$, use the mean $\mu(\mathbf{x}_n)$ and variance $\sigma(\mathbf{x}_n)^2$ of the resultant normal distribution as the approximate predictive mean and variance, and replace the integral with respect to $f_n$ with its point estimate. See [9] for further details.

## 4 Collaborative Multi-output Gaussian Processes for Novelty Detection

As the novelty detector, we follow the paradigm of Kandemir et al. [6] and propose the recent collaborative multi-output Gaussian process (CGP) regression model [8] as the feature predictor. The CGP is defined as

$$p(\mathbf{G}, \mathbf{U}|\mathbf{Z}, \mathbf{X}) = \prod_{j=1}^{Q} \mathcal{SGP}(\mathbf{g}_j, \mathbf{u}_j|\mathbf{Z}_j, \mathbf{X}), \;\; p(\mathbf{H}, \mathbf{V}|\mathbf{W}, \mathbf{X}) = \prod_{i=1}^{P} \mathcal{SGP}(\mathbf{h}_i, \mathbf{v}_i|\mathbf{W}_i, \mathbf{X}),$$

$$p(\mathbf{Y}|\mathbf{G}, \mathbf{H}) = \prod_{i=1}^{P} \prod_{n=1}^{N} \mathcal{N}\Big(y_{in}\Big| \sum_{j=1}^{Q} w_{ij} g_j + h_i, \beta^{-1}\Big), \tag{2}$$

where $\mathbf{X}$ is a $N \times D$ matrix having $D$ dimensional $N$ instances in its rows, and $\mathbf{Y}$ is a $N \times P$ matrix having the corresponding $P$ dimensional targets in its rows. Here, $\mathbf{g}_j$'s are sparse GPs shared across all output dimensions, and $\mathbf{h}_i$'s are sparse GPs specific to each output dimension. The likelihood in Equation 2 combines all these GPs linearly by the weights $w_{ij}$. The intractable posterior of this model is inferred by the efficient stochastic variational inference method. See [8] for further details. We use the source code provided by the authors [1].

Given the feature vector of a candidate at the previous frame $\mathbf{x}_n$, we plug its top 50 principal components into the CGP as input, and predict the top 5 principal components $\mathbf{y}_n$ of its features at the current frame as the output. As the degree of surprise, or Training Utility Value (TUV) in other terms, Kandemir et al. [6] propose using the squared distance between the mean of the predicted outputs $\boldsymbol{\mu}(\mathbf{x}_n) = [\mu_1(\mathbf{x}_n), \cdots, \mu_P(\mathbf{x}_n)]$ and the true observations $TUV(\mathbf{x}_n)_{MSE} = \|\boldsymbol{\mu}(\mathbf{x}_n) - \mathbf{y}_n\|_2^2$. We here propose to extend this measure by taking into account also the predictive variance which is shown to be useful in certain recognition tasks [11]. A principled way for this is to define a true distribution for the observed features $p_{true} = \mathcal{N}(\hat{\mathbf{y}_n}|\mathbf{y}_n, \epsilon\mathbf{I})$, with a very small $\epsilon$, constructing spikes at the observed locations of the feature space, and use the Kullback-Leibler divergence between $p_{true}$ and the predictive distribution $p_{pred} = \mathcal{N}(\hat{\mathbf{y}_n}|\boldsymbol{\mu}(\mathbf{x}_n), \boldsymbol{\Sigma}_n)$ as the degree of surprise, where $[\boldsymbol{\Sigma}_n]_{ii} = \sigma_i^2(\mathbf{x}_n)$ is the predictive variance for output dimension $i$, and $\hat{\mathbf{y}_n}$ is the predicted feature vector. The resultant training utility function becomes

$$TUV(\mathbf{x}_n)_{KL} = \frac{1}{2}\left( tr(\boldsymbol{\Sigma}_n^{-1}\epsilon\mathbf{I}) + (\boldsymbol{\mu}(\mathbf{x}_n) - \mathbf{y}_n)^T \boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\mu}(\mathbf{x}_n) - \mathbf{y}_n) - \log\frac{|\epsilon\mathbf{I}|}{|\boldsymbol{\Sigma}_n|}\right).$$

Note that when $\boldsymbol{\Sigma}_n$ is equal for all instances, $TUV(\mathbf{x}_n)_{MSE}$ and $TUV(\mathbf{x}_n)_{KL}$ give identical orderings.

[1] `https://github.com/trungngv/multiplegp`

### 4.1 Proposal Generation by Sampling

Given the TUVs assigned to each instance, the question of how to use these values to choose the order for annotating the instances follows. We propose defining an $N$ category multinomial distribution by assigning each instance a probability of being chosen $P(C = \mathbf{x}_n) = \dfrac{TUV(\mathbf{x}_n)}{\sum_{j=1}^{N} TUV(\mathbf{x}_j)}$. We determine the next instance to be annotated by taking a draw from this distribution. In multi-armed bandit formulation, each choice $C = \mathbf{x}_n$ can be viewed as an arm (a potential action), and $P(C = \mathbf{x}_n)$ the reward distribution. The most intuitive way of choosing the instances with the highest reward probability (i.e. largest TUV) corresponds to the *greedy* algorithm (exploitation), which is known to have linear regret. We have also observed it to give poor performance. Since almost all highest ranking instances are true events, choosing instances only from the top of the list causes class imbalance. Our approach, sampling from the reward distribution, corresponds to *probability matching*, which is a well-known technique for balancing exploration and exploitation.

## 5 Workflows

*Way 1: Novelty detection plus supervised learning.* As novelty detector we compare the following models: i) **CGP-KL**: CGP followed by $TUV(\mathbf{x}_n)_{KL}$, ii) **CGP-MSE**: CGP followed by $TUV(\mathbf{x}_n)_{MSE}$, iii) **OC-SVM**: The standard One-Class Support Vector Machine [10], iv) **150fps**: Sparse-coding based real-time novelty detection proposed by Lu et al. [7], as a representative of the dictionary learning based novelty detection methods.

The difference of **CGP-MSE** from [6] is that the former samples the instances to be annotated as in Section 4.1, and then trains a supervised classifier on the annotated instances, while the latter classifies the events directly by thresholding the TUVs.

*Way 2: Active learning.* We consider the following two active learning methods:

- **MES:** Maximum entropy sampling [12] annotates the instances in decreasing order with respect to: $TUV(\mathbf{x}_n) = \Big| p(y_n|\mathbf{X}_u, \mathbf{y}_u) - 0.5 \Big|$, where $\mathbf{X}_u$ is the set of instances for which the labels $\mathbf{y}_u$ are known, $p(y_n|\mathbf{X}_u, \mathbf{y}_u)$ the predictive distribution of any probabilistic classifier, and $\mathbf{x}_n$ is an instance in the unlabeled set. In the well-known *exploration-exploitation* trade-off, this technique relies only on exploitation and ignores exploring the feature space.
- **BALD:** Bayesian Active Learning by Disagreement [1] follows a principled Bayesian approach and quantifies the importance of an instance by the change it is expected to make in the entropy of the model posterior

$$TUV(\mathbf{x}_n) = \mathbb{H}[\boldsymbol{\theta}|\mathbf{X}_u, \mathbf{y}_u] - \mathbb{E}_{p(y_n|\theta, \mathbf{x}_n)}\Big[\mathbb{H}[\boldsymbol{\theta}|\mathbf{X}_u, \mathbf{y}_u, \mathbf{x}_n, y_n]\Big],$$

where $\mathbb{H}[\cdot] = \mathbb{E}_p(\cdot)[-\log p(\cdot)]$ stands for the entropy function and $p(\boldsymbol{\theta}|\mathbf{X}_u, \mathbf{y}_u)$ is the posterior of a model. BALD has been shown to handle the exploration-exploitation trade-off in a balanced way, since the first term always favors the most uncertain instance (i.e. equals to MES), hence, performs exploitation, and the second term performs exploration by penalizing the terms with high intrinsic noise. The BALD for the standard GP classifier follows as

$$TUV(\mathbf{x}_n) = \mathit{\Phi}\left(\frac{\mu(\mathbf{x}_n)}{\sqrt{\sigma(\mathbf{x}_n)^2+1}}\right) - \frac{\sqrt{\frac{\pi \log 2}{2}}}{\sqrt{\sigma(\mathbf{x}_n)^2 + \frac{\pi \log 2}{2}}} \exp\left(\frac{\mu(\mathbf{x}_n)^2}{2\sigma(\mathbf{x}_n)^2 + \pi \log 2}\right).$$

## 6 Experiments

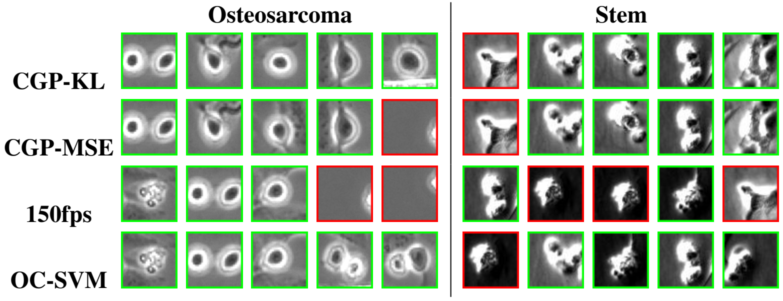
We detect mitosis and apoptosis events in phase-contrast microscopy sequences independently acquired from two different live cell tissues: i) a human osteosarcoma cell line that consists of eight sequences of 134 frames each, and ii) the public stem cell line data set of Huh et al. [4] that consists of three sequences of 540 frames each. For the stem cell data set, only apoptotic events were publicly available. We also annotated the mitotic events to make the two applications comparable. In all experiments, a model is trained on one sequence and tested on another. All results are averaged over all possible training-test combinations of the sequences of a given tissue type.

We used the standard Gaussian process classifier with a probit link likelihood as the supervised event predictor (the *Classifier* box in Figure 1) and with a squared exponential kernel function with isotropic covariance $k(\mathbf{x}, \mathbf{x}') = \gamma_0 \exp(-\gamma_1 \|\mathbf{x} - \mathbf{x}'\|_2^2)$. The kernel hyperparameters $\gamma_0$ and $\gamma_1$ are learned by Type II Maximum Likelihood. Since MES and BALD cannot perform an absolute cold-start (an empty labeled set), we started these models from a randomly chosen 10 labeled instances.

### 6.1 Event Detection from Few Annotations

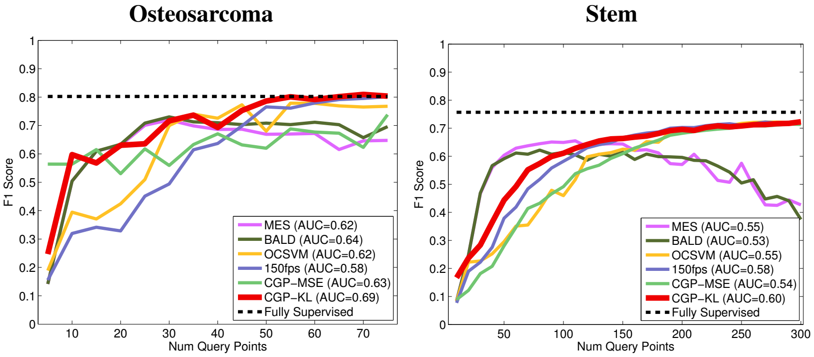
The main goal of our workflow is to direct the annotator to the relevant instances in the sequence. We achieve this without any supervision, differently from Kandemir et al.'s MOGP [6], which requires a small number of frames without any events to be provided. Figure 2 shows highest ranking five cell event candidates found by the four novelty detectors. CGP-KL, CGP-MSE, and OC-SVM all retrieve comparably relevant candidates. However, as seen in Figure 3, the TUVs found by CGP-KL lead to the steepest learning curve with respect to F1 score (harmonic mean of precision and recall), when a GP classifier is trained with an increasing number of annotated instances. For the novelty detectors, new training instances are chosen in each round using the sampling method described in Section 4.1. For the AL models, the unlabeled instances with largest TUV are chosen, following the theoretically grounded and commonly adopted way. While for the AL models the choice of new instances at a round is dependent on the choices and the trained classifiers of the previous rounds, the novelty detectors are trained once on an unsupervised sequence, and then provide a labeling order on a target sequence independently of the classifier. Despite not using any labeled data for proposal generation, the novelty detectors have a more stable learning curve. In both cell types, the two AL algorithms either saturate immaturely as for osteosarcoma, or overfit and harm the classifier in the later rounds as for the stem cell. For osteosarcoma, the fully supervised learning performance is exactly reached, and even slightly exceeded

**Fig. 2.** Highest ranking cell event candidates are sorted from left to right in decreasing TUV. Green frame around a patch indicates that the candidate corresponds to an event, and red frame indicates that it is a false positive.



after only $10(initial) + 15(rounds) \times 5(questions) = 85$ annotations (2.4% of the candidates), and for the stem cell, 95% of the fully supervised learning performance is covered after $10(initial) + 30(rounds) \times 10(questions)$ annotations (1.9% of the candidates).

**Fig. 3.** Learning curves of the sparse GP classifier when it is trained with instances chosen by the novelty detectors and the AL algorithms. A higher Area Under Learning Curve (AUC) indicates faster learning, which is desired. CGP-KL provides highest AUC in both tissue types.



In both tissue types, CGP-KL outperforms Kandemir et al. [6] with $F1 = 0.81$ vs. $F1 = 0.77$ for Osteosarcoma, and $F1 = 0.74$ vs. $F1 = 0.65$ for Stem from comparably few annotator effort ($< 3\%$ of candidates from both classes for CGP-KL versus five frames from the negative class for [6]). We observed that linear combination of the TUV's of CGP-KL and any of the AL algorithms never improves on CGP-KL alone.

Failure cases of our entire framework include floating dead cells, events taking place at the image boundaries or behind the white grid in osteosarcoma sequences, and events taking place more slowly than usual.

## 7 Discussion

The suboptimal performance of the AL models in the problem we studied is due to the severe class imbalance coming from the nature of the event detection problem, which makes the positive instances look overly valuable for the AL models. Hence, these models focus only on refining the decision boundary, ending up overfitting. This could also be seen from the fact that MES and BALD follow very similar learning patterns for both applications. In other words, BALD reduces to MES under class imbalance. Another reason for their suboptimal performance is the cold-start problem. AL algorithms require a labeled starting subset (warm start) that includes instances from both classes, which raises yet another novelty detection problem. On the contrary, the novelty detectors benefit from the class imbalance by using it as a modeling assumption.

The fact that clearly higher accuracies can be reached with a small annotation effort could be noteworthy for further studies in the image-based cell behavior analysis field. Our pipeline can be integrated into an annotation software used by biologists, and could provide them an importance ordering of the locations to be looked in large sequences. This would bring a remarkable effort gain given that expert annotator time is very costly.

## References

1. N. Houlsby, F. Huszar, Z. Ghahramani, and J.M. Hernández-Lobato. Collaborative Gaussian processes for preference learning. In *NIPS*, 2012.
2. S. Huh and M. Chen. Detection of mitosis within a stem cell population of high cell confluence in phase-contrast microscopy images. In *CVPR*, 2011.
3. S. Huh, Dai-Fei E. Ker, R. Bise, M. Chen, and Takeo Kanade. Automated mitosis detection of stem cell populations in phase-contrast microscopy images. *Trans. Medical Imaging*, 30(3):586–596, 2011.
4. S. Huh, H. Su, T. Kanade, et al. Apoptosis detection for adherent cell populations in time-lapse phase-contrast microscopy images. In *MICCAI*. 2012.
5. M. Kaakinen, S. Huttinen, L. Paavolainen, V. Marjomaki, J. Heikkila, and L. Eklund. Automatic detection and analysis of cell motility in phase-contrast time-lapse images using a combination of maximally stable extremal regions and kalman filter approaches. *Journal of Microscopy*, 253(1):65–78, 2014.
6. M. Kandemir, J. C. Rubio, U. Schmidt, J. Welbl, B. Ommer, and F. A. Hamprecht. Event Detection by Feature Unpredictability in Phase-Contrast Videos of Cell Cultures. In *MICCAI*, 2014.
7. C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in MATLAB. In *ICCV*, pages 2720–2727, 2013.
8. V.T. Nguyen and E. Bonilla. Collaborative multi-output Gaussian processes. In *UAI*, 2014.
9. C.E. Rasmussen and C.I. Williams. Gaussian processes for machine learning. 2006.
10. B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.
11. M. Seeger. Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations. *PhD Thesis*, 2003.
12. M.C. Shewry and H.P. Wynn. Maximum entropy sampling. *Journal of Applied Statistics*, 14(2):165–170, 1987.
13. E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *NIPS*, 2006.