Frame-Based Classification of Operation Phases in Cataract Surgery Videos

Manfred Jüergen Primus[1] ✉, Doris Putzgruber-Adamitsch[2] Mario Taschwer[1], Bernd Münzer[1], Yosuf El-Shabrawi[2], Laszlo Böszörmenyi[1], and Klaus Schoeffmann[1]

[1] Alpen-Adria Universität Klagenfurt, Austria
[juergen.primus|mt|bernd|laszlo|ks]@itec.aau.at
[2] Klinikum Klagenfurt am Wörthersee, Austria
[doris.putzgruber-adamitsch|yosuf.el-shabrawi]@kabeg.at

Abstract. Cataract surgeries are frequently performed to correct a lens opacification of the human eye, which usually appears in the course of aging. These surgeries are conducted with the help of a microscope and are typically recorded on video for later inspection and educational purposes. However, post-hoc visual analysis of video recordings is cumbersome and time-consuming for surgeons if there is no navigation support, such as bookmarks to specific operation phases. To prepare the way for an automatic detection of operation phases in cataract surgery videos, we investigate the effectiveness of a deep convolutional neural network (CNN) to automatically assign video frames to operation phases, which can be regarded as a single-label multi-class classification problem. In absence of public datasets of cataract surgery videos, we provide a dataset of 21 videos of standardized cataract surgeries and use it to train and evaluate our CNN classifier. Experimental results display a mean F1-score of about 68% for frame-based operation phase classification, which can be further improved to 75% when considering temporal information of video frames in the CNN architecture.

Keywords: medical multimedia, deep learning, video analysis, surgical workflow analysis

1 Introduction

Cataract is known as clouding of the eye's lens, a defect that often occurs in the course of aging. It affects the human visual system and has a tremendous negative impact on the patient's quality of life. This condition can be treated with a surgical procedure during which the natural lens is removed and an artificial lens is implanted, which usually results in a noticeable improvement of vision. Cataract surgery is by far the most frequently performed surgical procedure in the medical specialty of ophthalmology and one of the most frequently performed procedures across all specialties world-wide. It therefore follows broadly accepted common rules and can be called a *quasi-standardized procedure*. Cataract surgery

is usually performed in local anesthesia within 5 to 10 minutes, unless complications occur. High-volume surgeons usually operate several cataract surgeries within a single day. The surgeon looks at the patient's eye through an optical microscope for appropriate visualization and magnification. Furthermore, surgical microscopes usually have an additional optical system with a mounted camera to acquire a video signal. This video stream is displayed on a monitor and can also be recorded on a digital medium.

The facts that (1) a video signal is inherently available without any additional effort, (2) the course of action is well standardized, and (3) the procedure is frequently performed, make this specific domain an interesting subject of medical multimedia research. One of the fundamental problems in this field is automatic understanding of the surgical workflow and, in particular, temporal segmentation of a video into surgical phases. Such an automatic segmentation can greatly support surgeons in coping with their potentially huge video archives. It may even open the door for a comprehensive video documentation, which is not widely used yet due to the lack of video organization and navigation support. Beyond that, real-time processing methods may even support surgeons during the procedure in order to recognize or prevent adverse events. Such situations could be identified by detecting deviations from the surgical process model, causing the system to immediately alert the surgeon and provide context-sensitive assistance.

In this paper, we address the problem of differentiating between surgical phases in cataract surgery videos with a frame-based classification approach. Each video frame is classified separately as belonging to one of multiple operation phases. The task can therefore be considered as a multi-class single-label classification problem. The proposed frame-based classifier may be used to build an automatic operation phase detection system in future work. As underlying classification framework we use deep convolutional neural networks (CNN), which have proven to be very expedient for similar tasks with other types of surgery videos [5], but—to the best of our knowledge—have not been applied to ophthalmic surgery videos before. In addition to applying a CNN to a raw dataset of cataract surgery videos, we propose and evaluate two data preprocessing methods that aim at improving classification performance: (1) training data purification and balancing, and (2) adding temporal information of video frames. Since public datasets of cataract surgery videos are not yet available, we created such a dataset with ground-truth annotations to evaluate our proposed approach and provide it for public use by the scientific community.

Although acquiring videos from cataract surgeries and performing frame classification might seem to be a straightforward task, we encountered numerous challenges. They mainly relate to the strong domain specificity of ophthalmic videos, which require thorough analysis and adaptations of established techniques. For example, the visual appearance can vary considerably due to different individual preferences of surgeons regarding positioning of the microscope in terms of angle, zoom level, and light configuration. These settings can also change during a procedure. Moreover, the video camera that is mounted at the secondary optics is independent of the main optics and therefore needs to be

adapted separately by an assistant in case the surgeon changes settings. If this is not done properly, the video quality can be considerably impaired. Another major challenge in this domain is the necessity to incorporate highly skilled domain experts, i.e. experienced surgeons. Their knowledge about the specific characteristics and semantics of the videos is essential. However, as they only can spare a limited amount of time, it is crucial to provide appropriate annotation tools to extract their expert knowledge as efficiently as possible.

The contributions of this paper are: (1) We propose to apply convolutional neural networks (CNN) to frame-based operation phase classification of cataract surgery videos and obtain promising results; (2) we show that classification performance of our approach can be further improved by (a) dataset purification and balancing and (b) adding temporal information of video frames as input to the CNN; (3) we provide a novel public dataset containing video recordings of 21 cataract surgeries and corresponding ground-truth annotations in terms of operation phase boundaries.

2 Related Work

We focus on related work concerning image understanding techniques applied to recorded videos of surgeries. Video recordings in the medical domain can primarily be found in the context of endoscopic or microscopic surgery. Literature in this field is mostly concerned with classification of instruments, actions, anatomy, and surgical workflow.

Early methods focus on hand-crafted features and similarity measures to detect and recognize instruments used in endoscopy. Speidel et al. [9] used images captured by a stereo endoscope and segment the potential shaft region based on saturation, brightness and amount of reddish color. The tip of the instrument is segmented with the help of a Bayesian classifier, before the instrument is recognized based on the normalized contour and distances to 3D representations of each instrument.

A bag-of-visual-words representation of SIFT-, SURF-, and ORB-features was used to train a support vector machine (SVM) for recognizing instruments and operation phases in cholecystectomy surgeries (i.e. removal of the gallbladder) [6,7]. The authors improved their approach by segmenting the image area into parts where an instrument might show up and parts that only show tissue. For classification only the potential instrument area was used.

Petscharnig et al. [4] proposed the use of transfer learning based on the AlexNet CNN architecture for frame-based classification of actions and anatomy in gynecologic surgery videos. In follow-up work [5] they showed that a GoogLeNet CNN model trained from scratch outperformed their AlexNet-model as well as an SVM classifier using off-the-shelf AlexNet features. Their best performing network achieved an F1-score of 85%.

Twinanda et al. [11] trained a CNN called EndoNet based on the AlexNet architecture for the classification of operation phases of cholecystectomy surgeries. The authors concatenated the FC7 layer of the AlexNet architecture with

the subsequent output layer to a new fully connected layer. A refinement of classifications was achieved using a hierarchical hidden Markov model.

In the field of cataract surgeries, Lalys et al. [3] used visual information such as histograms, texture, and shape for the classification of surgical tasks using an SVM. The classified images were aligned to already annotated recordings using a hidden Markov model and dynamic time warping. Charriere et al. [1] used a Bayesian network and two conditional random fields for classification of operation phases in cataract surgery videos.

Quellec et al. [8] introduced a method that divides cataract surgeries into ten phases. Each phase is divided into an action phase—where the surgical task is performed—and an idle phase—where almost nothing happens in the operation area, because the next step is prepared out of the microscope's sight or instruments are exchanged. Recorded videos were used to learn the differences between action phases and idle phases. A conditional random field was used to align phases of new videos to existing ones.

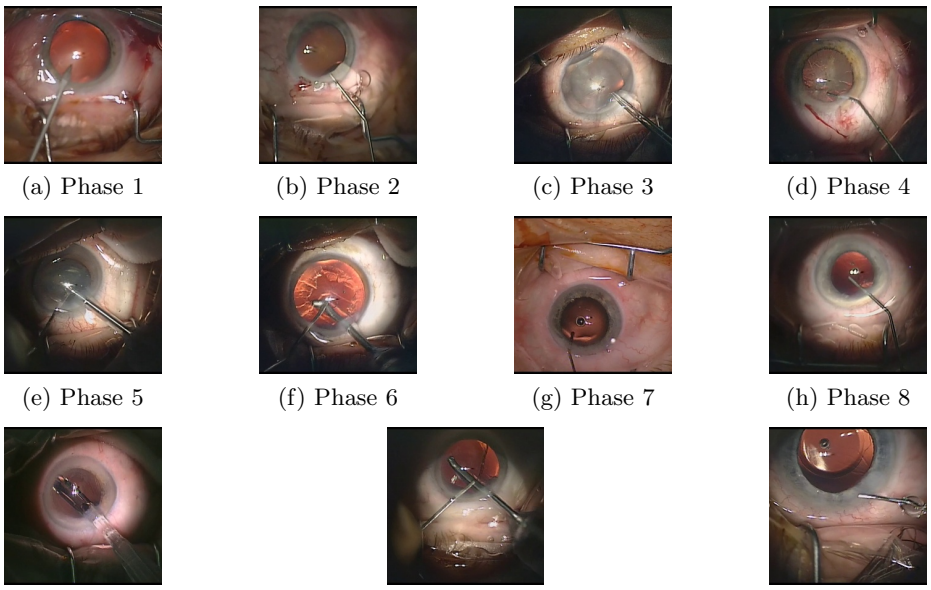
Previous multimedia research in the domain of cataract surgeries used the strict sequential order of the surgical workflow to detect phase transitions and did not yet employ CNNs. In contrast, our work addresses the classification of single frames of cataract surgery videos using newly trained CNNs. The major advantage of this approach is that it can be easily extended to additional classes pertaining to optional operation phases, out-of-order phases, or complications.

3 Cataract Surgery Dataset

Cataract surgery can be divided in eleven phases, which are: 1. Incision, 2. Viscous agent injection I, 3. Rhexis, 4. Hydrodissection, 5. Phacoemulsification, 6. Irrigation and aspiration, 7. Capsule polishing, 8. Viscous agent injection II, 9. Lens implant setting-up, 10. Viscous agent removal, 11. Tonifying and antibiotics. This is the standardized sequence of a cataract surgery without complications. Still, it can happen that some steps are repeated. E.g. incisions (done in the first phase) need to be widened for the implantation of the artificial lens. It can also be necessary to moisten the operation area in some cases. Representative keyframes of the standardized phases are shown in Figure 1.

The dataset, which we use for training and evaluation of CNN models, has been kindly provided by the ophthalmologic department of our medical partner. It consists of 21 single video recordings of cataract surgeries performed by four different surgeons and following closely the standardized surgical procedure. Videos containing optional phases (e.g. "moistening" to moisturize dry eyes, or "blue vision" to facilitate the *Rhexis* phase and avoid complications) are not considered in this work.

The videos are recorded using MPEG-2 with a resolution of 720×576 pixels. The bitrate is about 6 Mb/s with a framerate of 25 frames per second. The average length of a recording of a cataract surgery is 6 minutes and 52 seconds with a standard deviation of 2 minutes and 38 seconds. The videos contain also irrelevant parts before the first phase and after the last phase. This is due to



(a) Phase 1 (b) Phase 2 (c) Phase 3 (d) Phase 4

(e) Phase 5 (f) Phase 6 (g) Phase 7 (h) Phase 8

(i) Phase 9 (j) Phase 10 (k) Phase 11

Fig. 1. Example images for each phase of a cataract surgery

Table 1. Distribution of video frames in the cataract surgery video dataset

Nr.	Phase	absolute number of frames	relative number of frames (in percent)
1.	Incision	8,896	4.19
2.	Viscous agent injection I	5,303	2.50
3.	Rhexis	16,602	7.81
4.	Hydrodissection	11,998	5.65
5.	Phacoemulsification	67,293	31.67
6.	Irrigation and aspiration	28,704	13.51
7.	Capsule polishing	9,654	4.54
8.	Viscous agent injection II	4,953	2.33
9.	Lens implant setting-up	13,868	6.53
10.	Viscous agent removal	32,888	15.48
11.	Tonifying and antibiotics	12,328	5.80
	Total:	212,487	100.0

the fact that the recording is started some seconds before the surgery starts and stopped a few seconds after the operation has ended. These parts of the videos are not used for training and evaluation of our CNN models. All phases including the transitions have been annotated by a surgeon according to the standardized workflow model of cataract surgeries.

Table 1 shows the distribution of video frames per operation phase in the cataract surgery video dataset, which is extremely unbalanced due to different durations of operation phases. The longest phase (phase 5) takes approximately one third of the surgery's duration, whereas phases 2 and 8 represent the shortest phases. Since phases 2 and 8 are identical with respect to both visual appearance and semantics (viscous agent injection), they are treated as a single class for classification purposes. The numbers of frames in the merged phases 2 and 8 sum up to roughly 5% of all video frames. The smallest class is then represented by phase 1 (4.19%).

Figure 1 shows example images from each phase. Typically, the pupil is wide opened and the iris is very small. The lens can appear in a reddish color, if light is reflected directly from the choroid or rather gray in case of very dense cataract or a vitreous hemorrhage. The position of the lens is not always centralized, which may affect analysis methods in a negative manner. Instruments appear typically from the left or from the right or from both sides. A cataract surgery uses a small set of instruments. Moreover, most of the instruments are used for a specific action and they are therefore characteristic for a phase.

Every phase consists of alternating action and idle periods. This fact has been exploited by Quellec et al. [8] to segment a cataract surgery video based on the occurrence of idle phases. The surgeon uses the instrument(s) belonging to a certain phase in the action period. The instrument can still be visible in the idle period. It can also be changed to the instrument that belongs to the next phase during this idle period. In that case no instrument is visible. This behavior can not be handled directly by a CNN architecture. However, the appearance of an eye changes during a cataract surgery with respect to texture and color, which can be modeled by a CNN.

The anonymized dataset including annotations is available under `ftp://ftp-itec.aau.at/pub/datasets/ovid/cat-21/`.

4 Frame-Based Classification of Cataract Surgery Videos

We propose to apply a convolutional neural network (CNN) to classify frames of cataract surgery videos with respect to the operation phase (class) the video frame belongs to. In an effort to obtain more effective classification models, we consider three different preprocessing techniques applied to the training dataset, leading to different CNN models: (1) basic training dataset (no preprocessing), (2) manually filtered ("purified") and automatically balanced dataset, and (3) the purified and balanced dataset with additional temporal information of each video frame. Details of these preprocessing techniques and the corresponding experimental setup will be described in Sections 4.1, 4.2, and 4.3, respectively.

The CNN is trained from scratch based on the GoogLeNet architecture [10]. This neural network was designed and trained for the 1000-class ImageNet Challenge (ILSVRC) and used to classify everyday images. The GoogLeNet architecture has 27 layers consisting of 5 pooling layers and 22 layers with parameters that include a modular structure involving nine inception modules, each using

Table 2. Full videos are randomly sampled for either training or evaluation dataset.

Nr.	Phase	absolute number of frames		relative number of frames (in percent)	
		Train.	Eval.	Train.	Eval.
1.	Incision	6,642	2,254	3.78	6.09
2.+8.	Viscous agent injection I + II	8,522	1,734	4.86	4.69
3.	Rhexis	13,594	3,008	7.75	8.13
4.	Hydrodissection	9,570	2,428	5.45	6.56
5.	Phacoemulsification	54,679	12,614	31.16	34.09
6.	Irrigation and aspiration	26,000	2,704	14.82	7.31
7.	Capsule polishing	8,223	1,431	4.69	3.87
9.	Lens implant setting-up	11,787	2,081	6.72	5.62
10.	Viscous agent removal	27,498	5,390	15.67	14.57
11.	Tonifying and antibiotics	8,973	3,355	5.11	9.07
	Total:	175,488	36,999	100.0	100.0

1×1 convolutions for dimensionality reduction. The input of the network are 224×224×3 sized RGB images shifted by a mean image. The prediction of the 1000 classes is done using a linear layer with softmax loss. To adapt the network architecture for prediction of ten phases of cataract surgery, we decreased the number of output neurons to 10. Otherwise the GoogLeNet architecture remains unchanged.

Training of CNN models is performed using the CAFFE framework [2]. The video frames are fed into a Lightning Memory-Mapped Database (LMDB), which is used as input for the CNN. The solver uses *Adam* as gradient-based optimization method provided by CAFFE. As base learning rate we use 0.001; momentum 1 and momentum 2 are set to 0.9 and 0.999, respectively. The training batch size is set to 64 images.

4.1 Basic Cataract-Surgery-Phase CNN

For the basic cataract surgery phase CNN model we partitioned the dataset described in Section 3 into a training and an evaluation subset. Out of the 21 videos in the dataset, 17 videos were chosen randomly for training. The remaining four videos are used for the evaluation of the CNN model. To train this first CNN model, we use the annotated videos of the training set as they are and split the videos into phases according to annotations without further refinement.

Table 2 shows the distribution of video frames in the resulting training and evaluation datasets. Phases *incision*, *irrigation and aspiration*, and *tonifying and antibiotics* have a large difference in the relative number of frames between training and evaluation datasets. This can be explained by the large variation in the duration of these phases. For example, phase *irrigation and aspiration* has an average duration of 55 seconds with a standard deviation of 40 seconds. All four *irrigation and aspiration* phases occurring in videos of the evaluation dataset

Table 3. Purified training dataset of cataract surgery videos

Nr. Phase	absolute number of frames	relative number of frames (in percent)
1. Incision	3,279	2.49
2.+8. Viscous agent injection	3,780	2.87
3. Rhexis	11,772	8.95
4. Hydrodissection	5,970	4.54
5. Phacoemulsification	49,986	38.02
6. Irrigation and aspiration	22,049	16.77
7. Capsule polishing	5,307	4.04
9. Lens implant setting-up	5,541	4.21
10. Viscous agent removal	17,447	13.27
11. Tonifying and antibiotics	6,353	4.83
Total:	131.484	100.0

have a duration below the average duration of this phase. Similar observations can be made for the other deviations.

The 175,488 images of the training dataset are center-cropped to a square shape and downsized to 256×256 pixels. This preparation reduces the size of the LMDB and the training time. Furthermore, we shuffle images before writing them into the LMDB to avoid feeding the CNN with a group of similar pictures when the LMDB is read sequentially. At training time we use the data augmentation methods provided by the CAFFE framework to vary the input on every training iteration (epoch): random mirroring and random cropping to the required size of 224×224 pixels.

4.2 Purified and Balanced Cataract-Surgery-Phase CNN

The first preprocessing method of the training dataset consists of manual purification followed by automatic balancing of the dataset. Purification takes care to reduce the variation of data within each class (operation phase). Each phase is characterized by the presence of certain instruments. During a phase there are also short periods, where these instruments are not visible. During manual purification, we identify all frames where none of these instruments are visible, and remove them from the training dataset. Table 3 shows that in total 44,004 images have been removed from the original training dataset.

It is interesting that some phases are affected less than others from purification. Phases *incision*, *viscous agent injection*, and *lens implant setting-up* lose approximately half of their samples, which can be explained as follows. In phase *incision* the surgeon performs two incisions, which are done within seconds. Between them and the end of the phase no instrument is visible. After the viscous agent is injected in phase eight, the lens implant setting-up is prepared, resulting in several seconds when no instrument is visible. The main instrument in phase *lens implant setting-up* is the cartridge for the lens, which is visible for

approximately half of the phase. The least affected phase is *phacoemulsification*, where only 8.6% of the images are dropped, because during the whole phase the phacoeomulsification-tip is visible except at the end of the phase when instruments are changed.

Purification makes the training dataset even more unbalanced, especially if we compare phase *incision* with 2.5% of the training images and phase *phacoemulsification* with more than 38% of the training images. This unbalanced dataset strongly increases the likelihood to classify in favor of the majority class. To overcome this problem we apply either random sampling (to video frames of large classes) or three data augmentation techniques (to frames of small classes): (1) simple copying, (2) rotation, and (3) scaling.

In detail, we choose 12,000 as uniform sample size for each phase. This means that larger classes are randomly reduced and smaller classes are extended with randomly chosen data augmentation techniques. For example, the phase *phacoemulsification* is reduced to one quarter of the original size, whereas phase *incision* is extended four-fold with artificially modified images.

To ensure that each available image is used for training, all images of the classes are copied in a first step. As long as a class has too many samples, one sample is randomly chosen and deleted. If the class has too few samples, we choose one of three data augmentation methods randomly and apply it to one randomly chosen (unmodified) image of the class. These steps are repeated until a uniform distribution of 12,000 images per class is achieved.

For rotation and scaling we randomly choose values for the rotation angle and the scaling factor. The angle for the rotation is constrained to the range $[-10°, +10°]$. The rotation introduces an empty area in the image, which is eliminated by cropping the image to a maximum-sized square whose corners hit these areas. Finally, the rotated image is resized to 256×256 pixels again. For scaling we select a square with a randomly chosen length between 246 and 156 pixels. We center-crop the image to this square and scale the resulting image up to a size of 256×256 pixels.

4.3 Timestamp-Based Cataract-Surgery-Phase CNN

For the third CNN model we extend each of the video frames with time information: the ratio of the frame number and the total number of frames in the video (relative timestamp). To feed timestamps into the CNN, we add a fourth "color channel"[3] to each image that contains this time information. Figure 2 shows that for most of the phases the starting time is well distinguishable. Temporal information of video frames is therefore expected to improve the classification performance of the trained CCN model.

[3] This decision is due to restrictions of the CAFFE framework, which does not easily allow adding inputs to fully connected layers of the CNN.

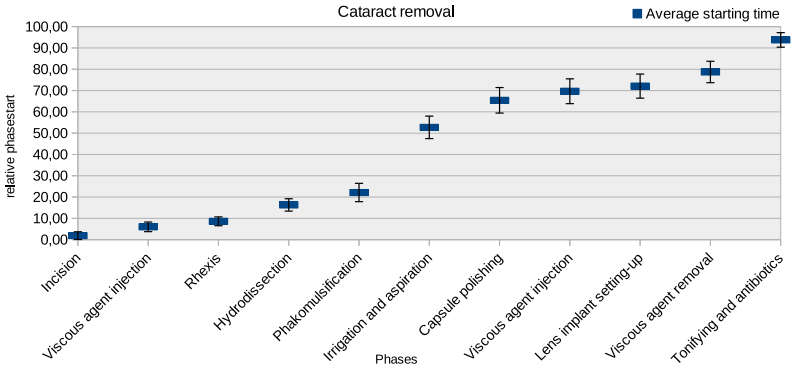


Fig. 2. Relative starting time of operation phases in surgery videos

Table 4. Classification results. Bold numbers indicate best performance within a phase.

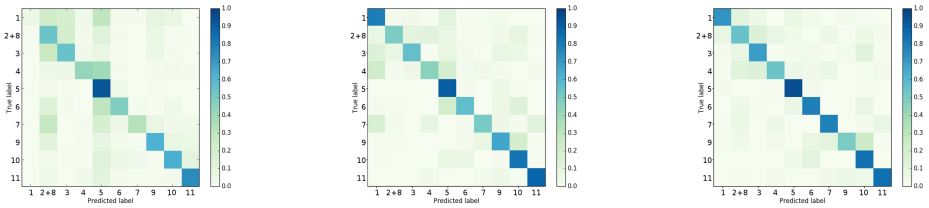
	Phase numbers										
CNN-Type	1	2+8	3	4	5	6	7	9	10	11	Average
Precision											
Basic CNN	**0.87**	0.24	0.61	0.70	0.73	0.55	0.61	0.62	**0.83**	0.76	0.65
Balanced CNN	0.38	**0.59**	**0.89**	0.66	0.89	**0.68**	**0.72**	0.39	0.75	0.90	0.69
Time-based CNN	0.65	0.37	0.82	**0.75**	**0.96**	**0.68**	0.71	**0.76**	0.75	**0.91**	**0.74**
Recall											
Basic CNN	0.06	0.54	0.55	0.43	0.93	0.48	0.33	0.62	0.63	0.75	0.53
Balanced CNN	**0.80**	0.49	0.56	0.45	0.91	0.57	0.51	**0.67**	0.83	**0.88**	0.67
Time-based CNN	0.72	**0.55**	**0.69**	**0.54**	**0.95**	**0.80**	**0.79**	0.50	**0.84**	0.85	**0.72**
F1-score											
Basic CNN	0.11	0.33	0.58	0.53	0.82	0.51	0.43	**0.62**	0.72	0.76	0.59
Balanced CNN	0.52	**0.54**	0.69	0.54	0.90	0.62	0.60	0.49	**0.79**	**0.89**	0.68
Time-based CNN	**0.69**	0.44	**0.75**	**0.62**	**0.95**	**0.73**	**0.75**	0.60	**0.79**	0.88	**0.73**

5 Evaluation

As evaluation dataset we use the four randomly selected videos mentioned in Section 4.1. It consists of 36,999 samples. The detailed distribution of video frames can be seen in Table 1.

For a given (preprocessed) training dataset, the CNN is trained for 50 epochs and the CNN model resulting from each epoch is kept for subsequent model selection. From the 50 resulting CNN models, only the best performing model (with respect to accuracy on the training dataset) is selected for final evaluation.

Table 4 structures the results in three quality measures: Precision, Recall, and F1-score (harmonic mean). In each of the table sections we see among each other the results of the three CNN-models: *Basic CNN* (Section 4.1), *balanced*



(a) Basic CNN (b) Balanced CNN (c) Time-based CNN

Fig. 3. Confusion matrices for the basic CNN, balanced CNN, and time-based CNN

CNN (Section 4.2), and *time-based CNN* (Section 4.3). It can be seen that each refinement improves the average performance of the network in terms of precision, recall, and F1-score clearly.

The *time-based CNN* shows problems with precision for phase 2+8 compared to other phases, where it performs similar or better than the other two CNN models. Figure 3c shows that *time-based CNN* tends to confuse phases 2 and 8 with neighboring phases 3 and 9, respectively.

A considerable performance gain can be achieved in terms of recall for the *balanced CNN* and the *time-based CNN*. Again, the *time-based CNN* outperforms both other networks for all phases but 1, 9, and 11, where it performs only slightly worse.

Looking at the F1-score we see that the *time-based CNN* shows a similar performance or outperforms the *balanced CNN* and the *basic CNN* in all cases except one (phase 2+8). In the absence of similar studies in the area of frame-based classification of ophthalmic surgery videos, we can compare our work only with results for frame-based classification of other types of surgery videos [4,5,6], where the authors achieve F1-scores of 0.51, 0.69, and 0.85.

Figure 3 visualizes the improvement in classification performance for *balanced CNN* and *time-based CNN* in comparison to *basic CNN*. The *basic CNN* has many false positive predictions for phase 5, which is overrepresented in the training dataset. There are also a lot of false positive predictions for phase 2+8. The confusion matrices also show that all CNN models have problems with the classification of phase 6.

6 Conclusion

In this paper, we examined frame-based classification of operation phases in cataract surgery videos using different CNN models. Along with this paper we provide a dataset of 21 video recordings of cataract surgeries that have been annotated by our medical partner. In particular we trained three CNN models based on the GoogLeNet-architecture. The basic CNN model was trained with a dataset that took the annotated phases directly as classes. For a second approach the dataset was modified manually by removing images where no instrument was

visible. This dataset was additionally balanced using different data augmentation techniques. Temporal information of video frames was added for training the third CNN model.

The evaluation showed that the classification performance can be improved significantly with a cleaned, balanced dataset and temporal information. In future work this CNN model can be extended for classification of various optional operation phases as well as for detection of complications. The development of such a neural network model enables also further automatic tools like keyframe selection for documentation, video summarization, or operation planning.

Acknowledgement. This work was supported by Universität Klagenfurt and Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214 U. 3520/26336/38165.

References

1. Charrière, K., Quellec, G., Lamard, M., Martiano, D., Cazuguel, G., Coatrieux, G., Cochener, B.: Real-time analysis of cataract surgery videos using statistical models. Multimedia Tools and Applications pp. 1–19 (2016)
2. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proc. of the 22nd ACM Int. Conf. on Multimedia. pp. 675–678. ACM (2014)
3. Lalys, F., Riffaud, L., Bouget, D., Jannin, P.: A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. IEEE Transactions on Biomedical Engineering 59(4), 966–976 (2012)
4. Petscharnig, S., Schöffmann, K.: Deep learning for shot classification in gynecologic surgery videos. In: Conf. Multimedia Modeling. pp. 702–713. Springer (2017)
5. Petscharnig, S., Schöffmann, K.: Learning laparoscopic video shot classification for gynecological surgery. Multimedia Tools and Applications pp. 1–19 (2017)
6. Primus, M.J., Schoeffmann, K., Böszörmenyi, L.: Instrument classification in laparoscopic videos. In: Content-Based Multimedia Indexing (CBMI), 2015, 13th International Workshop on. pp. 1–6. IEEE (2015)
7. Primus, M.J., Schoeffmann, K., Böszörmenyi, L.: Temporal segmentation of laparoscopic videos into surgical phases. In: Content-Based Multimedia Indexing (CBMI), 2016, 14th International Workshop on. pp. 1–6. IEEE (2016)
8. Quellec, G., Lamard, M., Cochener, B., Cazuguel, G.: Real-time segmentation and recognition of surgical tasks in cataract surgery videos. IEEE Transactions on Medical Imaging 33(12), 2352–2360 (2014)
9. Speidel, S., Benzko, J., Krappe, S., Sudra, G., Azad, P., Peter, B.: Automatic classification of minimally invasive instruments based on endoscopic image sequences. In: SPIE Medical Imaging. pp. 72610A–72610A (2009)
10. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conf. Computer Vision and Pattern Recognition. pp. 1–9 (2015)
11. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging 36(1), 86–97 (2017)