Supplementary: Large-scale, Fast and Accurate Shot Boundary Detection through Convolutional Neural Networks

I. Our data sets

Fig. 1 shows samples from the gradual class of our dataset (SBD_Syn). Our data is synthetically generated through image compositing. It is diverse, containing a wide variety of colors, texture, objects, motion and so on. Fig. 2 shows hard negative samples from our hard negative data (SBD_HN). The samples contain challenging cases that commonly confuse gradual transition detectors e.g. fast motion, fast zoom in, illumination changes, object occlusion, strong lighting, and so on. Fig. 3 shows 10 sequences from our synthetically generated wipe dataset. The sample shows that we use diverse alpha mats to generate our wipes data set.

Tab. I shows the significance and importance of our synthetic SBD_Syn and hard negative SBD_HN datasets in generating high quality detections. We evaluate our technique, DeepSBD, on different datasets with six different training sets: 1) R_3-5 2) R_3-6 3) R_3-6 + HN, 4) S + r, 5) S + r + HN and 6) and S + HN. S and HN is short for our datasets SBD_Syn and SBD_HN. R_3-6 represent TRECVID real videos and annotations from 2003 to 2006. r is T2005 and Baraldi. Results show that training with R_3-5 generate poor performance. In addition, it limits us to testing on just 3 data-sets. Adding T2006 to training improves performance but limits our testing further to 2 data-sets. Adding our hard negative data SBD_HN (HN) improves precision and performance significantly. This shows the high quality and importance of our SBD_HN. The best performance, however, is generated when both our datasets SBD_Syn and SBD_HN with r are used for training. In addition to the highest performance, this option allow us to test on all TRECVID videos, except T2005. Removing r from the training generates the second best performance. This, however, allow us to test on all TRECVID videos, including T2005. The experiment shows the significance and importance of our data-sets. We performed this experiment on several test sets and we found S + r + HN and S + HN are always the top and competitive to each other. This shows the significance of our datasets.

Tab. II-XV shows detailed per video results for different testing sets. For each testing dataset, we report the results using two different training-sets (S+r+HN and S+HN). We show: the number of transitions (#T), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

II. Processing speed

Fig. 4-5 examines the processing speed (test-phase) of our technique with different batch sizes as input. We ran our model on 6,394 segments. Each segment is 16-frame long, and hence our test-set contains 102,304 frames. Fig. 4 reports the total speed in seconds while Fig. 5 reports the real-time speed up factor. Tab. XVI shows detailed analysis of this experiment. For each batch size we ran our technique twice to ensure consistency. Results show that the processing speed gain from 10 to 100 batch size is not significant. Thats between 16-19.3 real-time speed up factor.

III. What does the network learn

Fig. 6 visualizes the feature response of our technique. We show the visualization of four different image sequences. For each sequence, we randomly selected two segments (16 frames) from UCF101 and synthetically generated a sharp and gradual transition using image compositing models. We treated one of the two sequences as no-transition. We examined all segments using our technique, DeepSBD. Fig. 6 shows the heat map of some Conv5 filter responses for each transition type. The filters are stacked next to each other, in blocks. The red grid shows some filters' borders. Time is the y-axis and space is the x-axis. Vertical space is averaged over the horizontal space. Sharp transitions have abrupt responses in the time axis in form of bright horizontal lines. Gradual transitions have blurred responses in the time axis. No transitions do not show a specific response pattern. The learned patterns of the three classes capture meaningful and discriminative information. Such information generate high detection results, as shown through out our results.

Fig. 1: 100 images from the gradual class of our dataset. Such data is generated synthetically through image composting.

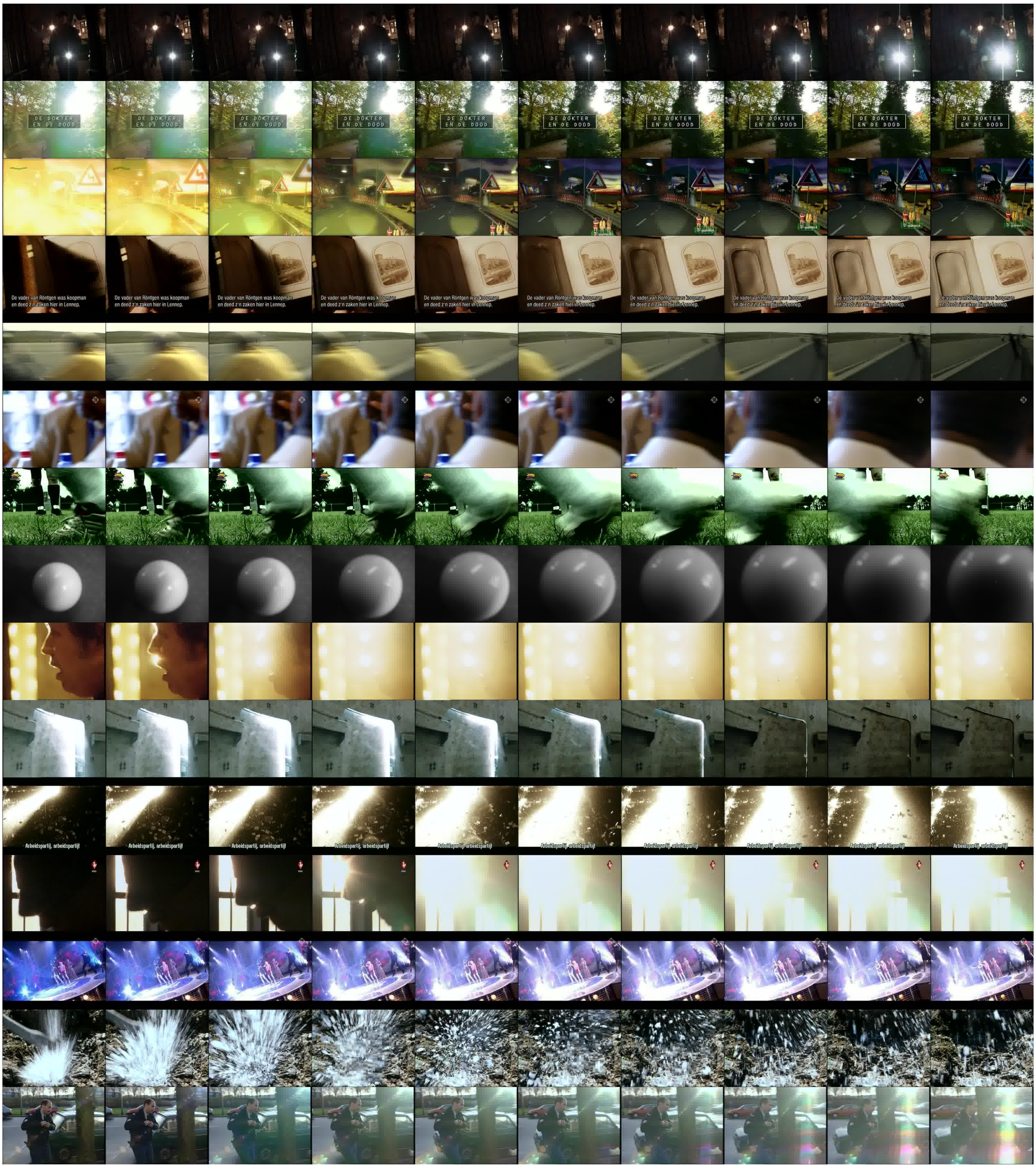


Fig. 2: Hard negative samples from our hard negative dataset. We carefully selected these samples through a semi-automated process. They represent complicated cases such as illumination variation, fast motion, occlusion and so on.

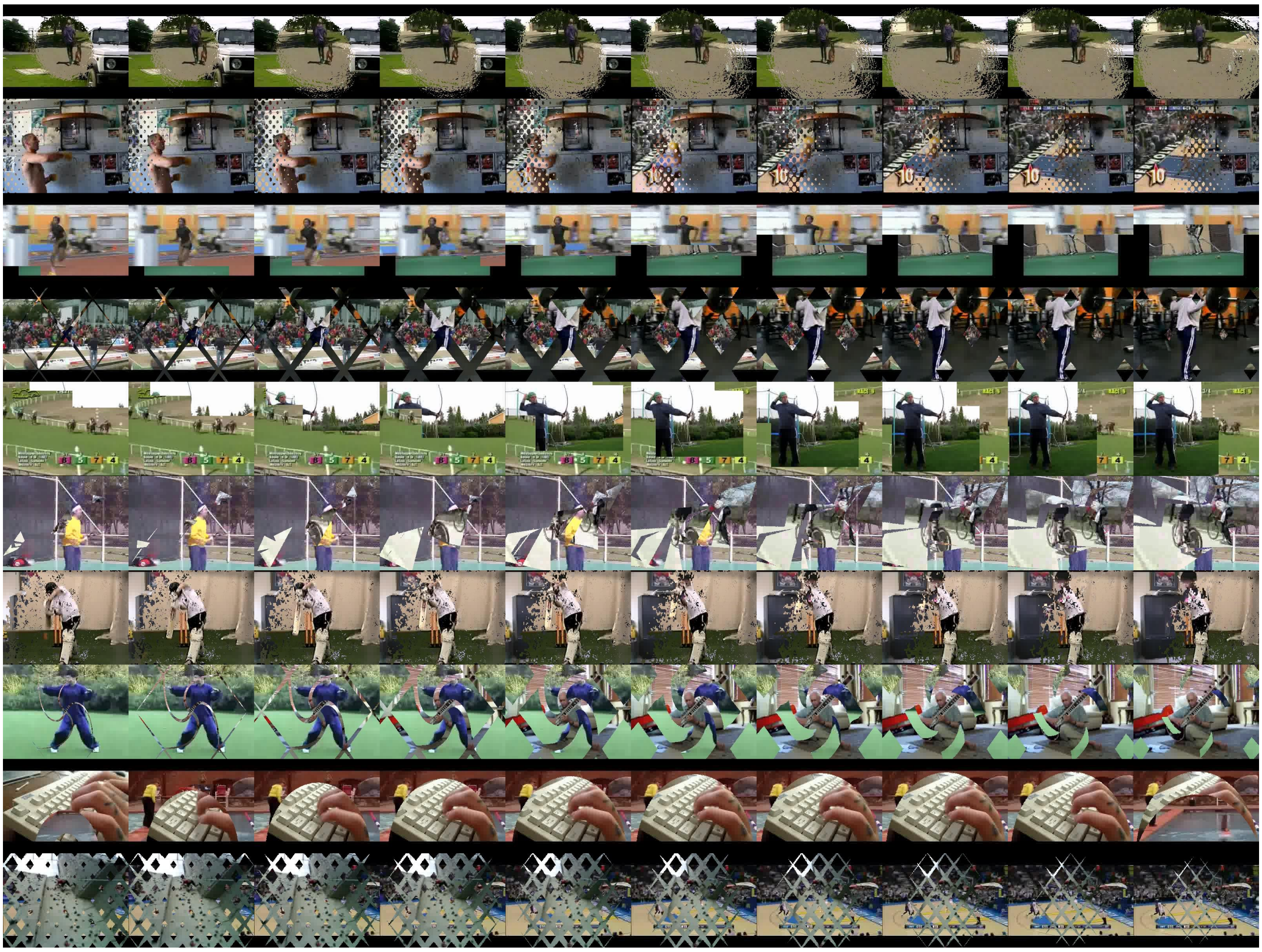


Fig. 3: Sample of 10 sequences from our synthetic wipe data set. We generate the wipes data set with considerably big amount of varying alpha mats.

TABLE I: Training our technique DeepSBD with different datasets. R_3-5 represent all TRECVID videos except 2001a, 2006 and 2007. Results show that the best performance is always generated when both our synthetic (S) and hard negative (HN) datasets are used (see S+r+HN and S+HN). Here, r is a very small portion of real videos (T2005 and Baraldi). The advantage of using S+HN is allowing us to test on all TRECVID videos, including T2005. Finally, our hard negative data HN clearly improves the precision and overall f-score.

	P	R	F	P	R	F
T2001a						
R_3-5	0.693	0.78	0.734	0.863	0.691	0.768
R_3-6	0.762	0.814	0.787	0.93	0.891	0.91
R_3-6+HN	0.917	0.753	0.827	0.96	0.923	**0.941**
S+r	0.782	0.851	0.815	0.926	0.92	0.923
S+r+HN	**0.951**	0.861	0.904	0.927	**0.936**	0.931
S+HN	0.934	**0.912**	**0.923**	**0.979**	0.904	**0.94**
T2006						
R_3-5	0.641	0.747	0.69	0.691	0.838	0.758
S+r	0.834	0.744	0.786	0.86	0.873	0.866
S+r+HN	**0.888**	0.804	**0.844**	0.863	**0.93**	**0.895**
S+HN	0.827	**0.834**	0.83	**0.876**	0.869	0.872
T2007						
R_3-5	0.495	0.665	0.568	0.894	0.872	0.883
R_3-6 +	0.683	0.683	0.683	0.957	0.95	0.953
R_3-6+HN	0.755	0.705	0.729	0.961	0.961	0.961
S+r	0.722	0.63	0.673	0.979	0.955	**0.967**
S+r+HN	**0.799**	**0.753**	**0.776**	**0.973**	**0.969**	**0.971**
S+HN	0.779	0.714	0.745	0.969	**0.966**	**0.968**
T2003						
S+r	0.735	0.703	0.718	**0.899**	0.837	0.867
S+r+HN	**0.779**	0.741	0.759	0.892	**0.842**	0.866
S+HN	0.741	**0.804**	**0.771**	**0.898**	**0.846**	**0.871**
T2004						
S+r	0.868	0.774	0.818	0.928	**0.929**	**0.929**
S+r+HN	**0.918**	0.819	0.866	0.923	**0.929**	**0.926**
S+HN	0.888	**0.884**	**0.886**	**0.941**	0.918	**0.929**
T2005						
S+HN	0.791	0.866	0.827	0.927	0.941	0.934

TABLE II: Detailed per video results of T2001b. Here, we use S+r+HN for training our model. We report the combined results for both gradual and sharp transitions. We show the true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual and Sharp					
Video	**TP**	**FP**	**FN**	**P**	**R**	**F**
BOR03	237	32	5	0.881	0.979	0.928
BOR08	456	8	75	0.983	0.859	0.917
BOR10	58	84	94	0.408	0.382	0.395
BOR12	117	5	19	0.959	0.86	0.907
BOR17	77	137	171	0.36	0.31	0.333
Total	945	266	364	0.78	0.722	0.75

TABLE III: Detailed per video results of T2001b. Here, we use S+HN for training our model. We report the combined results for both gradual and sharp transitions. We show the true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual and Sharp					
Video	**TP**	**FP**	**FN**	**P**	**R**	**F**
BOR03	240	30	2	0.889	0.992	0.938
BOR08	500	7	31	0.986	0.942	0.963
BOR10	54	82	98	0.397	0.355	0.375
BOR12	114	5	22	0.958	0.838	0.894
BOR17	66	106	182	0.384	0.266	0.314
Total	974	230	335	0.809	0.744	0.775

TABLE IV: Detailed per video results of T2002. Here, we use S+r+HN for training our model. We report the combined results for both gradual and sharp transitions. We show the true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual and Sharp					
Video	**TP**	**FP**	**FN**	**P**	**R**	**F**
01811a	60	7	4	0.896	0.938	0.916
6011	40	96	81	0.294	0.331	0.311
8024	85	22	21	0.794	0.802	0.798
8386	113	10	5	0.919	0.958	0.938
8401	26	5	5	0.839	0.839	0.839
10558a	122	1	8	0.992	0.938	0.964
23585a	149	10	16	0.937	0.903	0.92
23585b	103	3	1	0.972	0.99	0.981
34921a	70	4	5	0.946	0.933	0.94
34921b	91	10	8	0.901	0.919	0.91
36553	200	21	14	0.905	0.935	0.92
50009	44	28	14	0.611	0.759	0.677
50028	81	17	12	0.827	0.871	0.848
UGS01	164	8	12	0.953	0.932	0.943
UGS04	218	25	5	0.897	0.978	0.936
UGS05	21	6	9	0.778	0.7	0.737
UGS09	169	12	24	0.934	0.876	0.904
Total	1756	285	244	0.86	0.878	0.869

TABLE V: Detailed per video results of T2002. Here, we use S+HN for training our model. We report the combined results for both gradual and sharp transitions. We show the true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual and Sharp					
Video	**TP**	**FP**	**FN**	**P**	**R**	**F**
01811a	60	7	4	0.896	0.938	0.916
6011	39	96	82	0.289	0.322	0.305
8024	96	29	10	0.768	0.906	0.831
8386	114	5	4	0.958	0.966	0.962
8401	30	8	1	0.789	0.968	0.87
10558a	125	1	5	0.992	0.962	0.977
23585a	159	8	6	0.952	0.964	0.958
23585b	103	4	1	0.963	0.99	0.976
34921a	71	6	4	0.922	0.947	0.934
34921b	91	11	8	0.892	0.919	0.905
36553	202	26	12	0.886	0.944	0.914
50009	53	29	5	0.646	0.914	0.757
50028	89	18	4	0.832	0.957	0.89
UGS01	171	12	5	0.934	0.972	0.953
UGS04	222	15	1	0.937	0.996	0.965
UGS05	26	21	4	0.553	0.867	0.675
UGS09	176	17	17	0.912	0.912	0.912
Total	1827	313	173	0.854	0.913	0.883

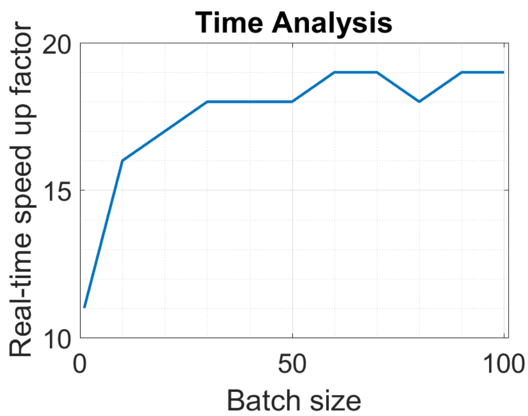


Fig. 5: Real-time speed factor of our technique. We report the results for different batch sizes as input.

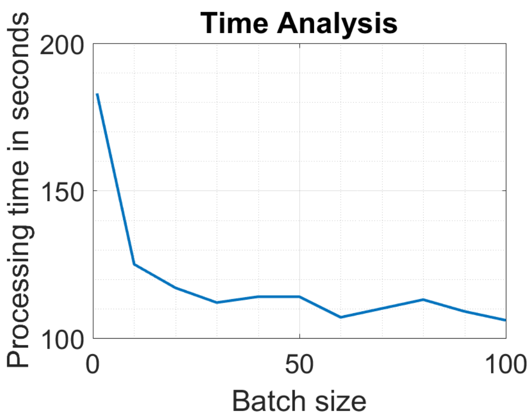


Fig. 4: Processing time in seconds of our technique. We report the results for different batch sizes as input.

TABLE VI: Detailed per video results of T2001. Here, we use S+r+HN for training our model. We report the results for both gradual and sharp transitions. For each class we show the number of transitions (#T), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual						**Sharp**						
Video	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**
BOR10_001	11	11	0	0	1	1	1	0	0	0	0	-	-	-
BOR10_002	11	9	0	2	1	0.818	0.9	0	0	0	0	-	-	-
NAD57	25	22	1	3	0.957	0.88	0.917	45	45	4	0	0.918	1	0.957
NAD58	44	37	0	7	1	0.841	0.914	40	33	0	7	1	0.825	0.904
anni001	8	6	0	2	1	0.75	0.857	0	0	1	0	0	-	-
anni005	27	26	2	1	0.929	0.963	0.945	39	36	13	3	0.735	0.923	0.818
anni006	31	27	3	4	0.9	0.871	0.885	42	41	0	1	1	0.976	0.988
anni007	5	4	0	1	1	0.8	0.889	5	5	0	0	1	1	1
anni008	13	12	0	1	1	0.923	0.96	2	2	0	0	1	1	1
anni009	64	57	3	7	0.95	0.891	0.919	40	37	0	3	1	0.925	0.961
anni010	56	50	11	6	0.82	0.893	0.855	98	84	1	14	0.988	0.857	0.918
Total	295	261	20	34	0.929	0.885	0.906	311	283	19	28	0.937	0.91	0.923

TABLE VII: Detailed per video results of T2001. Here, we use S+HN for training our model. We report the results for both gradual and sharp transitions. For each class we show the number of transitions (#T), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual						**Sharp**						
Video	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**
BOR10_001	11	11	0	0	1	1	1	0	0	0	0	-	-	-
BOR10_002	11	10	0	1	1	0.909	0.952	0	0	0	0	-	-	-
NAD57	25	21	2	4	0.913	0.84	0.875	45	45	1	0	0.978	1	0.989
NAD58	44	39	0	5	1	0.886	0.94	40	35	0	5	1	0.875	0.933
anni001	8	6	0	2	1	0.75	0.857	0	0	0	0	-	-	-
anni005	27	27	1	0	0.964	1	0.982	39	35	5	4	0.875	0.897	0.886
anni006	31	27	5	4	0.844	0.871	0.857	42	39	0	3	1	0.929	0.963
anni007	5	5	0	0	1	1	1	5	5	0	0	1	1	1
anni008	13	13	0	0	1	1	1	2	2	0	0	1	1	1
anni009	64	60	2	4	0.968	0.938	0.952	40	36	0	4	1	0.9	0.947
anni010	56	50	9	6	0.847	0.893	0.87	98	84	0	14	1	0.857	0.923
Total	295	269	19	26	0.934	0.912	0.923	311	281	6	30	0.979	0.904	0.94

TABLE VIII: Detailed per video results of T2003. Here, we use S+r+HN for training our model. We report the results for both gradual and sharp transitions. For each class we show the number of transitions (#T), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual						**Sharp**						
Video	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**
203_CNN	171	134	44	37	0.753	0.784	0.768	280	228	13	52	0.946	0.814	0.875
222_CNN	101	74	5	27	0.937	0.733	0.822	309	273	11	36	0.961	0.883	0.921
224_ABC	131	108	10	23	0.915	0.824	0.867	296	281	13	15	0.956	0.949	0.953
412_ABC	137	115	6	22	0.95	0.839	0.891	345	323	17	22	0.95	0.936	0.943
425_ABC	180	161	12	19	0.931	0.894	0.912	295	266	11	29	0.96	0.902	0.93
515_CNN	131	89	11	42	0.89	0.679	0.771	283	265	17	18	0.94	0.936	0.938
531_CNN	108	75	12	33	0.862	0.694	0.769	359	316	13	43	0.96	0.88	0.919
619_ABC	127	46	125	81	0.269	0.362	0.309	321	154	155	167	0.498	0.48	0.489
Total	1086	802	225	284	0.781	0.738	0.759	2488	2106	250	382	0.894	0.846	0.87

TABLE IX: Detailed per video results of T2003. Here, we use S+HN for training our model. We report the results for both gradual and sharp transitions. For each class we show the number of transitions (#T), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual						**Sharp**							
Video	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**
203_CNN	171	143	57	28	0.715	0.836	0.771	280	230	9	50	0.962	0.821	0.886
222_CNN	101	80	24	21	0.769	0.792	0.78	309	275	11	34	0.962	0.89	0.924
224_ABC	131	116	14	15	0.892	0.885	0.889	296	282	8	14	0.972	0.953	0.962
412_ABC	137	122	11	15	0.917	0.891	0.904	345	323	11	22	0.967	0.936	0.951
425_ABC	180	170	28	10	0.859	0.944	0.899	295	265	12	30	0.957	0.898	0.927
515_CNN	131	105	16	26	0.868	0.802	0.833	283	259	15	24	0.945	0.915	0.93
531_CNN	108	85	24	23	0.78	0.787	0.783	359	316	18	43	0.946	0.88	0.912
619_ABC	127	52	131	75	0.284	0.409	0.335	321	154	155	167	0.498	0.48	0.489
Total	1086	873	305	213	0.741	0.804	0.771	2488	2104	239	384	0.898	0.846	0.871

TABLE X: Detailed per video results of T2004. Here, we use S+r+HN for training our model. We report the results for both gradual and sharp transitions. For each class we show the number of transitions (#T), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual						**Sharp**							
Video	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**
1004_ABCa	203	166	13	37	0.927	0.818	0.869	224	213	22	11	0.906	0.951	0.928
1012_CNNa	170	136	13	34	0.913	0.8	0.853	215	194	15	21	0.928	0.902	0.915
1016_CNNa	150	119	9	31	0.93	0.793	0.856	242	214	13	28	0.943	0.884	0.913
1021_ABCa	175	154	13	21	0.922	0.88	0.901	240	230	18	10	0.927	0.958	0.943
1101_CNNa	204	172	20	32	0.896	0.843	0.869	191	187	11	4	0.944	0.979	0.961
1109_ABCa	170	151	10	19	0.938	0.888	0.912	257	246	15	11	0.943	0.957	0.95
1123_CNNa	126	93	29	33	0.762	0.738	0.75	236	214	10	22	0.955	0.907	0.93
1126_ABCa	189	168	12	21	0.933	0.889	0.911	273	261	23	12	0.919	0.956	0.937
1208_CNNa	137	112	15	25	0.882	0.818	0.848	212	196	17	16	0.92	0.925	0.922
1210_ABCa	159	140	8	19	0.946	0.881	0.912	271	252	14	19	0.947	0.93	0.939
1216_CNNa	153	119	11	34	0.915	0.778	0.841	197	187	26	10	0.878	0.949	0.912
1221_ABCa	195	149	14	46	0.914	0.764	0.832	217	197	27	20	0.879	0.908	0.893
Total	2031	1679	167	352	0.91	0.827	0.866	2031	2591	211	184	0.925	0.934	0.929

TABLE XI: Detailed per video results of T2004. Here, we use S+HN for training our model. We report the results for both gradual and sharp transitions. For each class we show the number of transitions (#T), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual						**Sharp**							
Video	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**
1004_ABCa	203	177	9	26	0.952	0.872	0.91	224	209	18	15	0.921	0.933	0.927
1012_CNNa	170	149	23	21	0.866	0.876	0.871	215	191	13	24	0.936	0.888	0.912
1016_CNNa	150	122	13	28	0.904	0.813	0.856	242	211	12	31	0.946	0.872	0.908
1021_ABCa	175	154	22	21	0.875	0.88	0.877	240	227	14	13	0.942	0.946	0.944
1101_CNNa	204	187	13	17	0.935	0.917	0.926	191	180	12	11	0.938	0.942	0.94
1109_ABCa	170	159	12	11	0.93	0.935	0.933	257	241	11	16	0.956	0.938	0.947
1123_CNNa	126	99	32	27	0.756	0.786	0.77	236	206	8	30	0.963	0.873	0.916
1126_ABCa	189	179	16	10	0.918	0.947	0.932	273	260	14	13	0.949	0.952	0.951
1208_CNNa	137	117	22	20	0.842	0.854	0.848	212	192	17	20	0.919	0.906	0.912
1210_ABCa	159	148	21	11	0.876	0.931	0.902	271	251	7	20	0.973	0.926	0.949
1216_CNNa	153	137	25	16	0.846	0.895	0.87	197	184	21	13	0.898	0.934	0.915
1221_ABCa	195	168	18	27	0.903	0.862	0.882	217	195	13	22	0.938	0.899	0.918
Total	2031	1796	226	235	0.888	0.884	0.886	2775	2547	160	228	0.941	0.918	0.929

TABLE XII: Detailed per video results of T2006. Here, we use S+r+HN for training our model. We report the results for both gradual and sharp transitions. For each class we show the number of transitions (#T), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual						**Sharp**							
Video	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**
LNA	198	147	11	51	0.93	0.742	0.826	45	31	26	14	0.544	0.689	0.608
NFC	77	57	11	20	0.838	0.74	0.786	121	115	8	6	0.935	0.95	0.943
NEC	94	88	3	6	0.967	0.936	0.951	74	57	5	17	0.919	0.77	0.838
HNA	124	107	4	17	0.964	0.863	0.911	24	21	8	3	0.724	0.875	0.792
3PGC	228	171	32	57	0.842	0.75	0.794	132	112	48	20	0.7	0.848	0.767
CLE	123	98	22	25	0.817	0.797	0.807	244	236	33	8	0.877	0.967	0.92
CDC	302	231	27	71	0.895	0.765	0.825	139	129	65	10	0.665	0.928	0.775
8NNE	214	184	44	30	0.807	0.86	0.833	424	418	36	6	0.921	0.986	0.952
CLE	37	28	7	9	0.8	0.757	0.778	57	54	7	3	0.885	0.947	0.915
5PGC	190	155	44	35	0.779	0.816	0.797	81	75	30	6	0.714	0.926	0.806
MNE	181	156	11	25	0.934	0.862	0.897	339	323	21	16	0.939	0.953	0.946
CLE	27	25	4	2	0.862	0.926	0.893	44	42	0	2	1	0.955	0.977
1NNE	146	134	11	12	0.924	0.918	0.921	120	118	5	2	0.959	0.983	0.971
Total	1941	1581	231	360	0.873	0.815	0.843	1844	1731	292	113	0.856	0.939	0.895

TABLE XIII: Detailed per video results of T2006. Here, we use S+HN for training our model. We report the results for both gradual and sharp transitions. For each class we show the number of transitions (#T), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual						**Sharp**							
Video	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**
LNA	198	150	16	48	0.904	0.758	0.824	45	39	31	6	0.557	0.867	0.678
NFC	77	56	22	21	0.718	0.727	0.723	121	115	8	6	0.935	0.95	0.943
NEC	94	81	8	13	0.91	0.862	0.885	74	46	6	28	0.885	0.622	0.73
HNA	124	113	33	11	0.774	0.911	0.837	24	22	6	2	0.786	0.917	0.846
3PGC	228	168	38	60	0.816	0.737	0.774	132	105	41	27	0.719	0.795	0.755
CLE	123	110	28	13	0.797	0.894	0.843	244	223	14	21	0.941	0.914	0.927
CDC	302	241	40	61	0.858	0.798	0.827	139	119	53	20	0.692	0.856	0.765
8NNE	214	183	53	31	0.775	0.855	0.813	424	372	28	52	0.93	0.877	0.903
CLE	37	35	7	2	0.833	0.946	0.886	57	51	3	6	0.944	0.895	0.919
5PGC	190	149	42	41	0.78	0.784	0.782	81	72	17	9	0.809	0.889	0.847
MNE	181	168	26	13	0.866	0.928	0.896	339	294	17	45	0.945	0.867	0.905
CLE	27	26	5	1	0.839	0.963	0.897	44	28	0	16	1	0.636	0.778
1NNE	146	138	21	8	0.868	0.945	0.905	120	116	3	4	0.975	0.967	0.971
Total	1941	1618	339	323	0.827	0.834	0.83	1844	1602	227	242	0.876	0.869	0.872

TABLE XIV: Detailed per video results of T2007. Here, we use S+r+HN for training our model. We report the results for both gradual and sharp transitions. For each class we show the number of transitions (#T), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

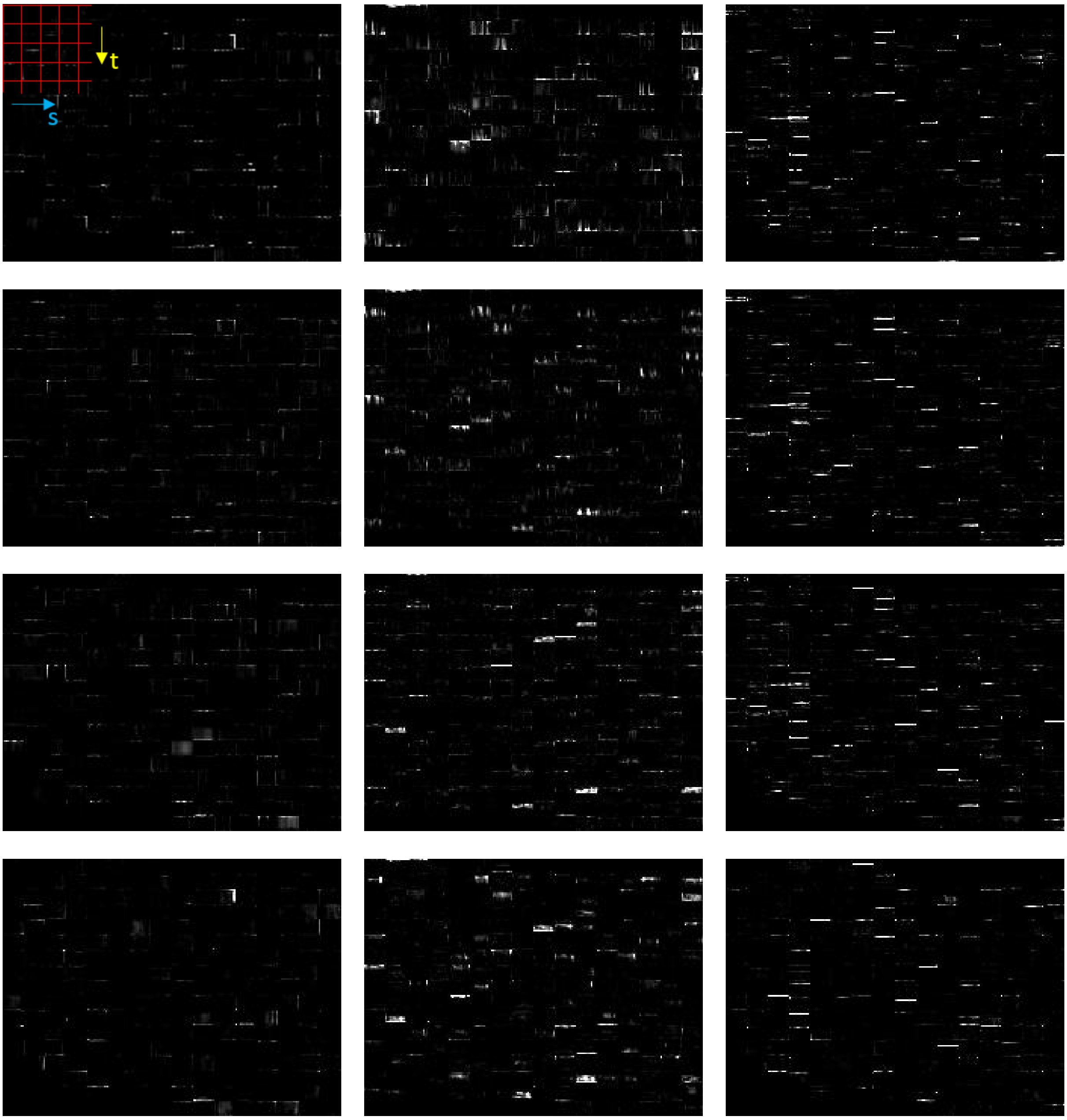
	Gradual						**Sharp**							
Video	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**
BG_11362	4	2	2	2	0.5	0.5	0.5	104	95	14	9	0.872	0.913	0.892
BG_14213	61	49	0	12	1	0.803	0.891	106	106	3	0	0.972	1	0.986
BG_2408	20	17	4	3	0.81	0.85	0.829	101	100	5	1	0.952	0.99	0.971
BG_34901	16	9	4	7	0.692	0.562	0.621	224	215	4	9	0.982	0.96	0.971
BG_35050	4	1	0	3	1	0.25	0.4	98	98	0	0	1	1	1
BG_35187	23	19	3	4	0.864	0.826	0.844	135	125	2	10	0.984	0.926	0.954
BG_36028	0	0	0	0	-	-	-	87	86	9	1	0.905	0.989	0.945
BG_36182	14	3	0	11	1	0.214	0.353	95	95	1	0	0.99	1	0.995
BG_36506	6	4	1	2	0.8	0.667	0.727	77	76	0	1	1	0.987	0.993
BG_36537	30	24	13	6	0.649	0.8	0.716	259	243	0	16	1	0.938	0.968
BG_36628	10	5	3	5	0.625	0.5	0.556	192	187	2	5	0.989	0.974	0.982
BG_37359	6	6	1	0	0.857	1	0.923	164	158	1	6	0.994	0.963	0.978
BG_37417	12	9	2	3	0.818	0.75	0.783	76	73	2	3	0.973	0.961	0.967
BG_37822	10	9	1	1	0.9	0.9	0.9	119	115	3	4	0.975	0.966	0.97
BG_37879	4	2	1	2	0.667	0.5	0.571	95	91	0	4	1	0.958	0.978
BG_38150	4	4	0	0	1	1	1	215	213	2	2	0.991	0.991	0.991
BG_9401	3	3	0	0	1	1	1	89	88	0	1	1	0.989	0.994
Total	227	166	35	61	0.826	0.731	0.776	227	2164	48	72	0.978	0.968	0.973

TABLE XV: Detailed per video results of T2007. Here, we use S+HN for training our model. We report the results for both gradual and sharp transitions. For each class we show the number of transitions (#T), true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and F-measure (F).

	Gradual						**Sharp**							
Video	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**	**#T**	**TP**	**FP**	**FN**	**P**	**R**	**F**
BG_11362	4	0	2	4	0	0	-	104	81	13	23	0.862	0.779	0.818
BG_14213	61	45	2	16	0.957	0.738	0.833	106	106	3	0	0.972	1	0.986
BG_2408	20	18	5	2	0.783	0.9	0.837	101	100	7	1	0.935	0.99	0.962
BG_34901	16	8	2	8	0.8	0.5	0.615	224	219	6	5	0.973	0.978	0.976
BG_35050	4	0	1	4	0	0	-	98	98	0	0	1	1	1
BG_35187	23	19	1	4	0.95	0.826	0.884	135	125	3	10	0.977	0.926	0.951
BG_36028	0	0	0	0	-	-	-	87	86	14	1	0.86	0.989	0.92
BG_36182	14	3	0	11	1	0.214	0.353	95	95	3	0	0.969	1	0.984
BG_36506	6	4	2	2	0.667	0.667	0.667	77	76	1	1	0.987	0.987	0.987
BG_36537	30	24	20	6	0.545	0.8	0.649	259	244	0	15	1	0.942	0.97
BG_36628	10	6	3	4	0.667	0.6	0.632	192	191	5	1	0.974	0.995	0.985
BG_37359	6	6	1	0	0.857	1	0.923	164	157	3	7	0.981	0.957	0.969
BG_37417	12	10	2	2	0.833	0.833	0.833	76	72	4	4	0.947	0.947	0.947
BG_37822	10	9	1	1	0.9	0.9	0.9	119	115	5	4	0.958	0.966	0.962
BG_37879	4	3	1	1	0.75	0.75	0.75	95	92	0	3	1	0.968	0.984
BG_38150	4	4	2	0	0.667	1	0.8	215	214	1	1	0.995	0.995	0.995
BG_9401	3	3	1	0	0.75	1	0.857	89	89	0	0	1	1	1
Total	227	162	46	65	0.779	0.714	0.745	2236	2160	68	76	0.969	0.966	0.968

TABLE XVI: The processing time of our technique. We report detailed analysis of different batch sizes as input. The bigger the batch size, the less processing time is required. This, however, requires more GPU memory. Experiments shows that the processing speed gain from 10 to 100 batch size is not significant. Thats between 16-19.3 real-time speed up factor.

Batch size	Starting Time	End Time	# Seconds	Memory	# Iterations	Faster than real time by
1	19:08:23	19:11:26	183	69413912	6394	11.18076503
1	16:13:03	16:16:06	184	69413912	6394	11.12
10	21:16:31	21:18:37	125	694139048	640	16.36864
10	21:20:45	21:22:53	128	694139048	640	15.985
20	14:55:16	14:57:13	118	1388278088	320	17.33966102
20	14:59:21	15:01:19	117	1388278088	320	17.48786325
30	21:06:39	21:08:30	112	2082417128	214	18.26857143
30	21:10:44	21:12:36	112	2082417128	214	18.26857143
40	15:04:59	15:06:52	114	2776556168	160	17.94807018
40	15:09:55	15:11:56	120	2776556168	160	17.05066667
50	11:00:04	11:01:59	115	3470695208	128	17.792
50	14:50:18	14:52:11	114	3470695208	128	17.94807018
60	21:25:47	21:27:34	107	4164834248	107	19.12224299
60	16:07:51	16:09:44	113	4164834248	107	18.10690265
70	15:18:27	15:20:19	112	4858973288	92	18.26857143
70	15:22:29	15:24:20	110	4858973288	92	18.60072727
80	10:49:16	10:51:09	113	5553112328	80	18.10690265
80	10:54:25	10:56:19	114	5553112328	80	17.94807018
90	15:50:29	15:52:19	109	6247251368	72	18.77137615
90	15:55:18	15:57:08	111	6247251368	72	18.43315315
100	21:32:57	21:34:43	106	6941390408	64	19.30264151
100	21:36:45	21:38:32	106	6941390408	64	19.30264151



(a) No transition (b) Gradual transition (c) Sharp transition

Fig. 6: Filter responses of our technique DeepSBD stacked next to each other. The red grid shows some filters' borders. Here, y-axis is time and x-axis is space. Sharp transitions (c) have an abrupt response in time (bright horizontal lines). Gradual transitions (b) have blurred responses in time. No transition (a) do not show specific patterns.