Temporal Binary Representation for Event-Based Action Recognition

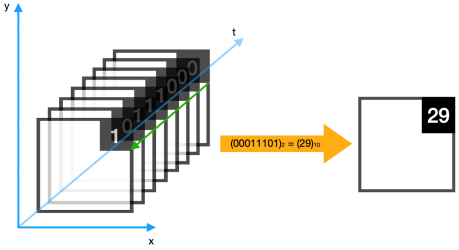
Simone Undri Innocenti, Federico Becattini, Federico Pernici, Alberto Del Bimbo
{name.lastname}@unifi.it

Introduction

- Event cameras capture illumination changes at extremely fast rates, generating an asynchronous stream of polarized events for each pixel.
- In order to use standard frame-based machine learning approaches, such as Convolutional Neural Networks, events must be aggregated into synchronous frames.
- Most event aggragation strategies lead to a loss of information by temporally quantizing the signal.
- We propose Temporal Binary Representation, a memory efficient event aggregation strategy, lossless up to a configurable temporal scale.

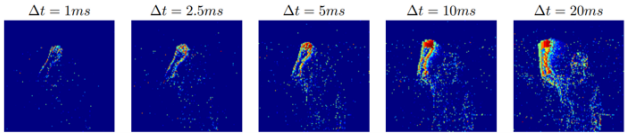
Temporal Binary Representation

- Given an arbitrarily small accumulation time Δt we build an intermediate binary representation b^i by checking the presence or absence of an event for each pixel.
- Stacking N temporally consecutive binary representations, each pixel can be considered as a binary string of N digits $[b^0_{x,y}, b^1_{x,y}, ..., b^{N-1}_{x,y}]$.
- We convert the binary string into a decimal number and normalize it dividing it by N.



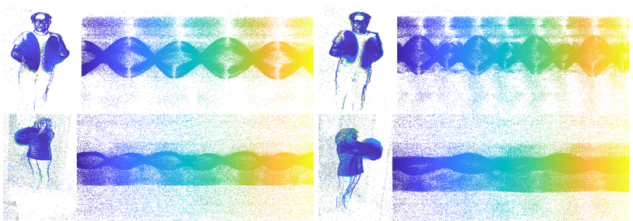
Properties

- Each frame covers a timespan of $N \times \Delta t$
- Memory efficient: N separate representations are encoded into a single frame preserving all information - less data to be processed by a Neural Network.
- Movement direction directly encoded in the image: no need to encode polarity since recent events have higher values.
- Lossless representation up to Δt which can be chosen arbitrarily small.



Action Recognition

- We evaluate our approach on the DVS128 Gesture Dataset by training two different models: Inception 3D and AlexNet + LSTM
- We collect the MICC-EVENT Gesture Dataset to increase the variability of DVS128: 640x480 resolution, multiple speed, different scales, different camera and orientations, uneven illumination.



Results

- DVS128 Gesture Dataset

	10 classes	11 classes
Time-surfaces [25]	96.59	90.62
SNN eRBP[26]	-	92.70
Slayer [27]	-	93.64
CNN [6]	96.49	94.59
Space-time clouds [28]	97.08	95.32
DECOLLE [29]	-	95.54
Spatiotemporal filt. [3]	-	97.75
RG-CNN [30]	-	97.20
Ours - AlexNet+LSTM	97.50	97.73
Ours - Inception3D	**99.58**	**99.62**

- MICC-EVENT Gesture Dataset

	TBR (ours)	Polarity	SAE
DVS128 Gesture Dataset	**99.62**	98.86	98.11
MICC-Event Gesture Dataset	**73.16**	68.40	70.13

MICC
Media Integration and Communication Center

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE