MFT: Long-Term Tracking of Every Pixel

Michal Neoral, Jonáš Šerých, Jiří Matas

Center for Machine Perception, Czech Technical University in Prague, FEE

The Dense Long-Term Tracking Task

Applications like structure-from-motion or video editing benefit from long-term dense correspondences. Optical Flow gives dense correspondences, but only between pairs of frames.

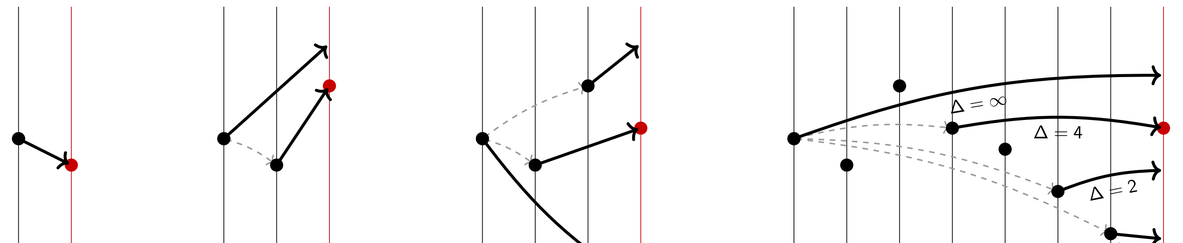
MFT: From two-frame optical flow to dense long-term trajectories

input: video. output (for every frame): current position + visibility for each point from the first frame

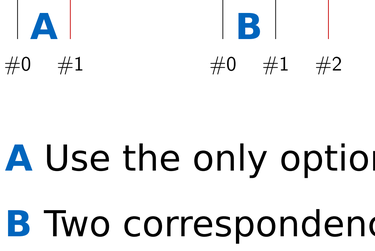
Baseline: Simple Chain of Flows

Cannot recover from temporary occlusions
Errors accumulate → drifting

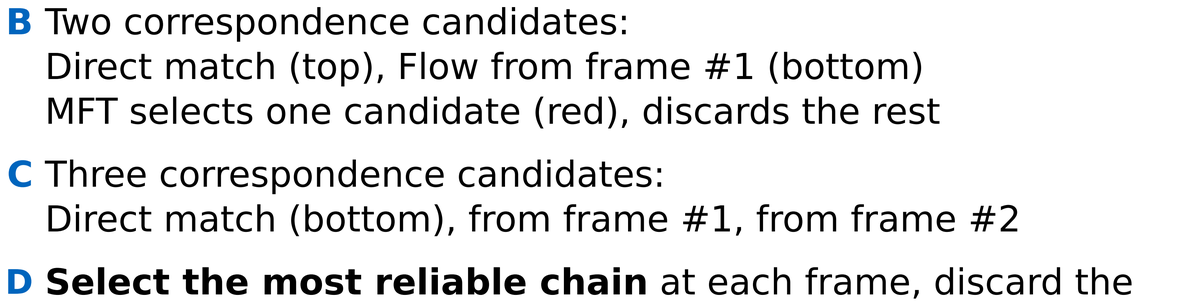
x-axis: frame number
y-axis: point position (1D for clarity)



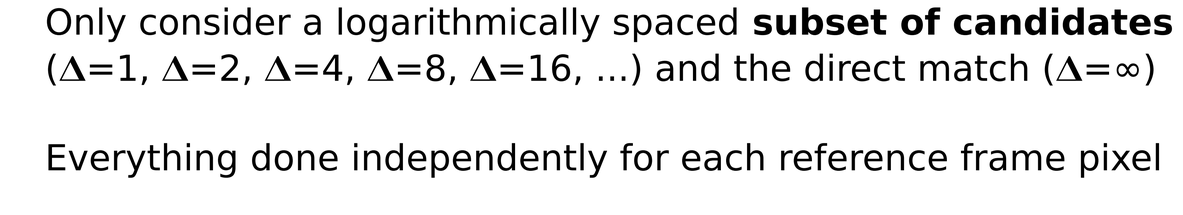
Baseline: Direct Flow From Template



Does not accumulate errors, no drift
Can recover after failure
But harder task - change of viewpoint, illumination, large motion



MFT: Select the Most Reliable Chain



A Use the only option: Optical flow from template to frame #1

B Two correspondence candidates:
Direct match (top), Flow from frame #1 (bottom)
MFT selects one candidate (red), discards the rest

C Three correspondence candidates:
Direct match (bottom), from frame #1, from frame #2

D **Select the most reliable chain** at each frame, discard the other candidates.

Only consider a logarithmically spaced **subset of candidates** (Δ=1, Δ=2, Δ=4, Δ=8, Δ=16, ...) and the direct match (Δ=∞)

Everything done independently for each reference frame pixel

Uncertainty Estimation

Select the candidate chain with the lowest uncertainty.
Estimate flow vector uncertainty with a small CNN head, chain the uncertainties by summation.

Trained on synthetic data using a standard uncertainty loss.

$$\mathcal{L}_u = \frac{1}{2\sigma^2} l_H(||\vec{x} - \vec{x}^*||_2) + \frac{1}{2}\log(\sigma^2)$$

flow error — *uncertainty score*

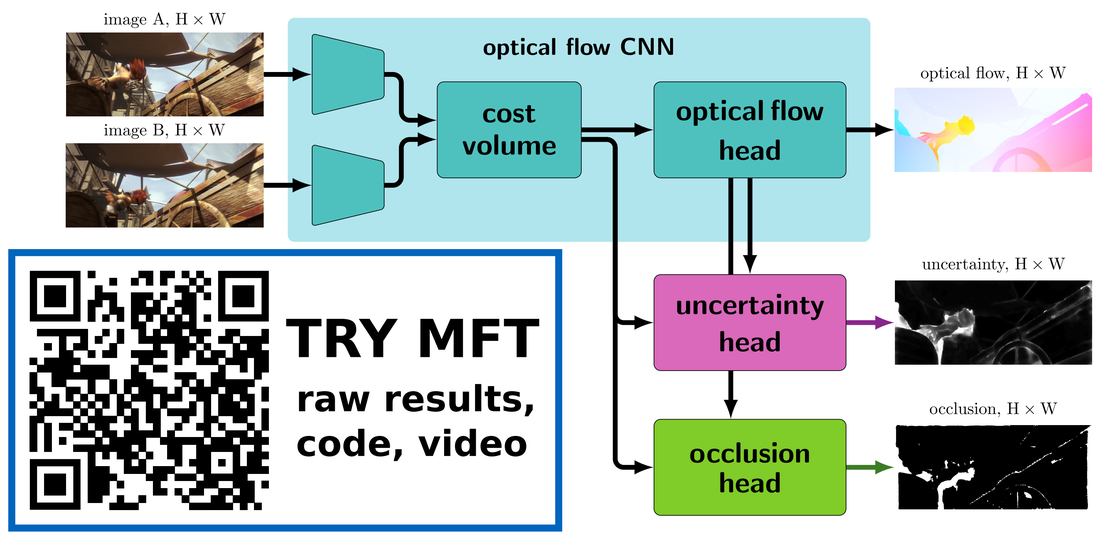
Occlusion Estimation

But do not select occluded chains.
Optical flow methods trained to work in occluded regions (using context).

Red arrow: flow correct, low uncertainty, but occlusion.
Continued tracking would switch from the head to the bamboo.

Trained on synthetic data using binary cross-entropy.

Optical Flow Adapted for MFT



Example MFT Application: Video Editing

A WOW! logo inserted in frame 0, color propagated by MFT to all frames

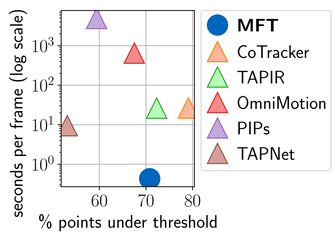
Point-Tracking Benchmark Results

	dense speed		*sparse tracking quality*								
			DAVIS - first			DAVIS - strided			Kinetics - first		
Method	sec / frame	FPS	AJ↑	$<\delta^x_{avg}$↑	OA↑	AJ↑	$<\delta^x_{avg}$↑	OA↑	AJ↑	$<\delta^x_{avg}$↑	OA↑
TAP-Net	9	0.11	33.0	48.6	78.8	38.4	53.1	82.3	38.4	54.4	80.6
PIPs	5000	0.0002	-	-	-	42.0	59.4	82.1	31.7	53.7	72.9
•OmniMotion	500	0.002	-	-	-	51.7	67.5	85.3	-	-	-
MFT (ours)	**0.4**	**2.32**	47.3	66.8	77.8	56.1	70.8	86.9	39.6	60.4	72.7
•TAPIR	25	0.04	56.2	70.0	86.5	61.3	72.3	87.6	49.6	64.2	85.0
•CoTracker	25	0.04	60.6	75.4	89.3	64.8	79.1	88.7	48.7	64.3	86.5

•*concurrent work* *FPS: speed of tracking a 512x512 video densely*

No dense benchmark → evaluation on sparse point-tracking.
TAP-Vid standard benchmark: ~20 GT point tracks per video

100+ FPS with pre-computed optical flows



MFT: good tracking quality dense trajectories at high speed

Acknowledgments

This work was supported by Toyota Motor Europe, by the Grant Agency of the Czech Technical University in Prague, grant No.SGS23/173/OHK3/3T/13, and by the Research Center for Informatics project CZ.02.1.01/0.0/0.0/16_019/0000765 funded by OP VVV.